Time Series Forecasting Project Report

Nandha Keshore Utti
PG-DSBA Online
March’ 22
Date: 22/10/2022

Contents

List of figures

List of tables

Problem Statement:

The data of different types of wine sales in the 20th century is to be analysed. Both of the given data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Sparkling Wines' Sales

1. Read the data as an appropriate Time Series data and plot the data.

Sample data frame as appropriate Time Series (TS) is as shown below:

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Table. 01

- Data has been loaded as TS.
- Sales data is given Month-wise.

Plot of TS data:

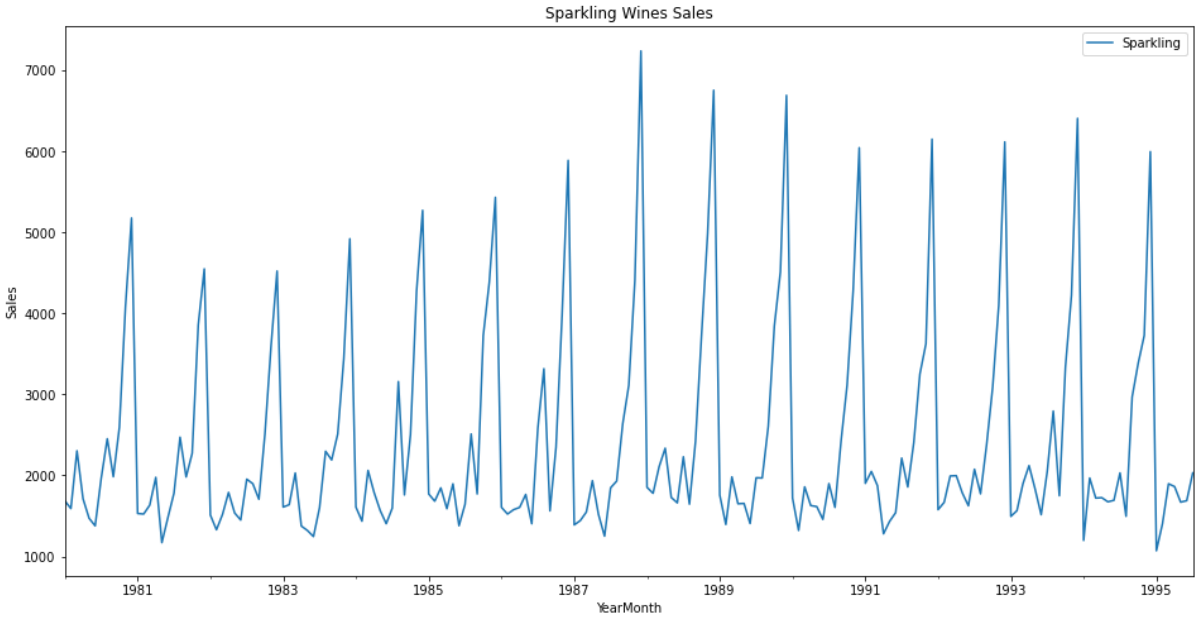

Fig. 01

Interpretation:

- There is no proper trend in the TS data.
- Seasonality is present in the TS data.

2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

EDA:

First five records of TS data are as shown below:

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Table. 02

Last five records of TS data are as shown below:

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Table. 03

- Sales data is given from Jan-1980 to July-1995, i.e., total 187 months is given.
- Maximum sales are occurring in December month of every year.

Data description: (for 187 months)

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

Table. 04

- Mean sales are ~2402.
- Maximum sales are 7242 (in Dec-1987).
- Minimum sales are 1070 (in Jan-1995).
- There are no null values.

- There are 11 duplicated records as shown below.

YearMonth	Sparkling
1984-01-01	1609
1985-09-01	1771
1986-02-01	1523
1987-05-01	1518
1988-08-01	1645
1990-08-01	1605
1992-01-01	1577
1994-02-01	1968
1994-03-01	1720
1995-03-01	1897
1995-07-01	2031

Table. 05

- All of the duplicated records are significant.

Decomposition:

- Let us perform decomposition of TS using additive method

Decomposition plot is as shown below:

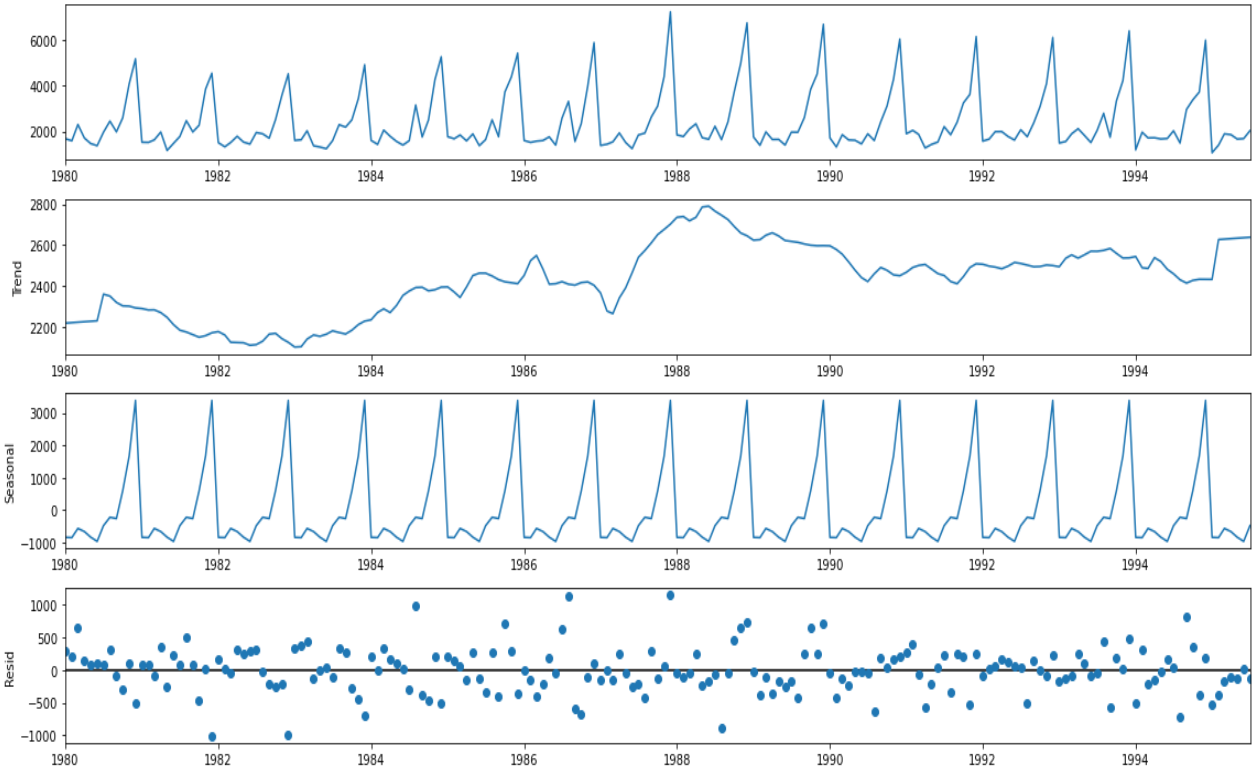

Fig. 02

Decomposition trend is as shown below:

```
YearMonth
1980-01-01    2219.309920
1980-02-01    2221.565688
1980-03-01    2223.821455
1980-04-01    2226.077222
1980-05-01    2228.332989
                  ...
1995-03-01    2629.859560
1995-04-01    2632.115327
1995-05-01    2634.371094
1995-06-01    2636.626861
1995-07-01    2638.882629
Name: trend, Length: 187, dtype: float64
```

Table. 06

Decomposition residuals are as shown below:

```
YearMonth
1980-01-01    304.056384
1980-02-01    215.091318
1980-03-01    642.372730
1980-04-01    145.453621
1980-05-01     74.727221
                  ...
1995-03-01   -170.665374
1995-04-01   -110.584483
1995-05-01   -132.310884
1995-06-01     13.605945
1995-07-01   -130.316468
Name: resid, Length: 187, dtype: float64
```

Table. 07

Decomposition seasonality is as shown below:

```
YearMonth
1980-01-01   -837.366304
1980-02-01   -845.657006
1980-03-01   -562.194185
1980-04-01   -659.530844
1980-05-01   -832.060210
                 ...
1995-03-01   -562.194185
1995-04-01   -659.530844
1995-05-01   -832.060210
1995-06-01   -962.232806
1995-07-01   -477.566161
Name: seasonal, Length: 187, dtype: float64
```

Table. 08

Evaluating the TS Decomposition: (for July-1995)

- Sales by TS decomposition are 2031.01
- Actual sales are 2031.

- TS decomposition of Sparkling wines sales by additive method working well.

3. Split the data into training and test. The test data should start in 1991.

Data split:

- Train set has 132 records and Test set has 55 records i.e., in ~70:30 ratio.
- Sample Train data is as shown below:

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Table. 09

- Sample Test data is as shown below:

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Table. 10

- We can see that Test data set is taken from Jan-1991.

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

1) Linear Regression Model:

- Let us fit the Linear Regression Model on Training data set.

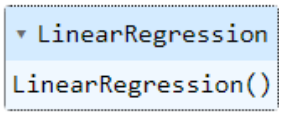

Fig. 03

- Plot showing Train, Test, Linear Regression Model best fit line is as shown below:

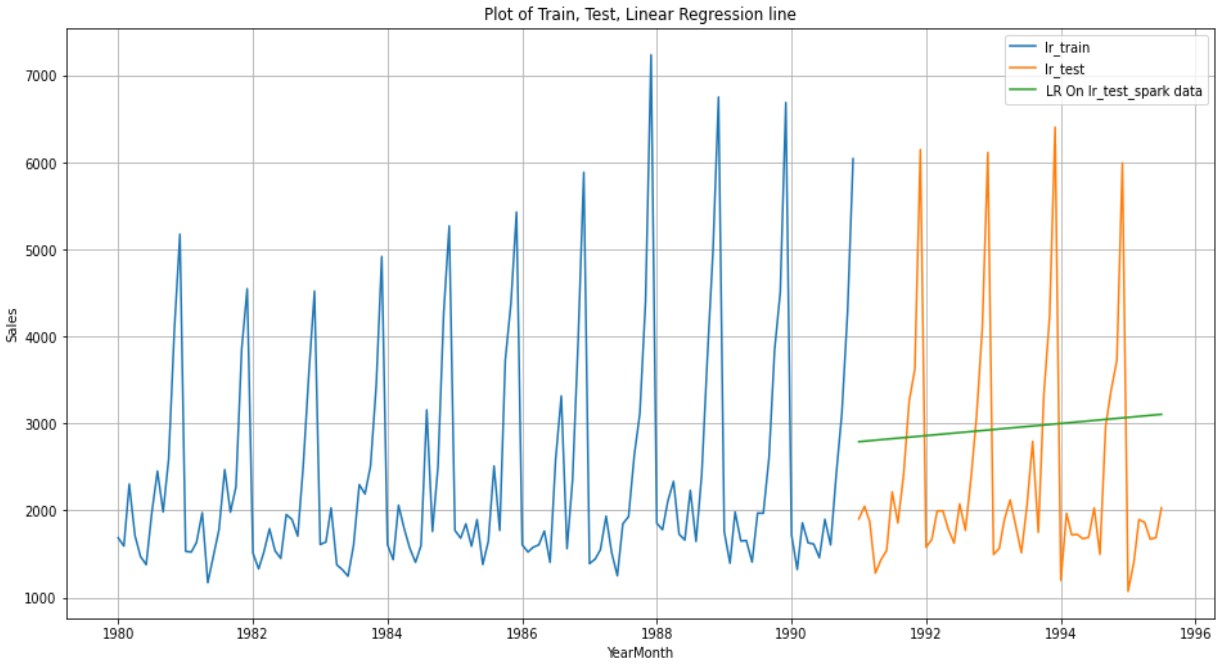

Fig. 04

- We can observe that Linear Regression model line and Test set curves are not matching.
- RMSE value is ~1389.

2) Naïve Approach:

- Let us use the last value of the Training data set to predict the test set.
- Last value of the Training data set 6047.
- Plot showing Train, Test, Naïve forecast line is as shown below:

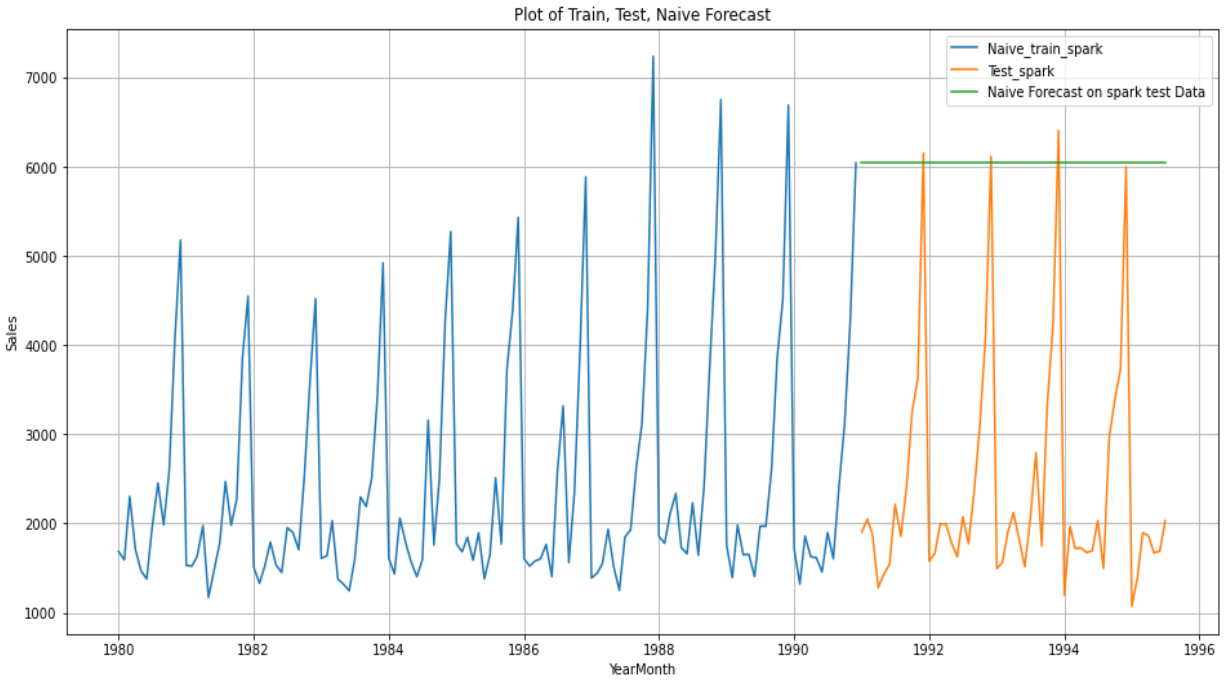

Fig. 05

- We can observe that Naïve forecast line and Test set curves are not matching.
- RMSE value is ~3864.

3) Simple Average Model:

- Let us use the Average sales' value of the Test data set to predict the test set.
- Average sales' value of the Test data set is ~2403.
- Plot showing Train, Test, Simple average line is as shown below:

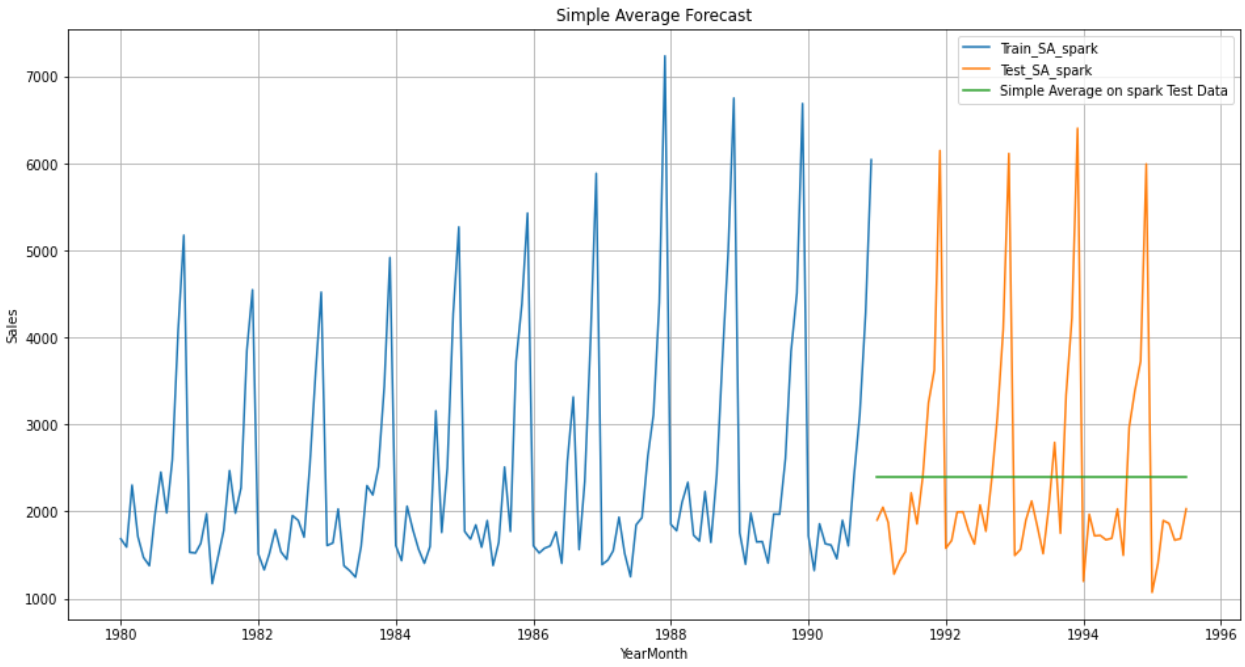

Fig. 06

- We can observe that Simple average forecast line and Test set curves are not matching.
- RMSE value is ~1275.

4) Moving Average Model:

- Let us use the 2, 4, 6, 8 Rolling mean sales' values of the Train data set to predict the Test set.
- Sample data frame showing 2, 4, 6, 8 Moving Average sales' values of the Training data set is shown below:

YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_8
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN

Table. 11

- Plot showing Train, Test, Moving average curves of Train set is as shown below:

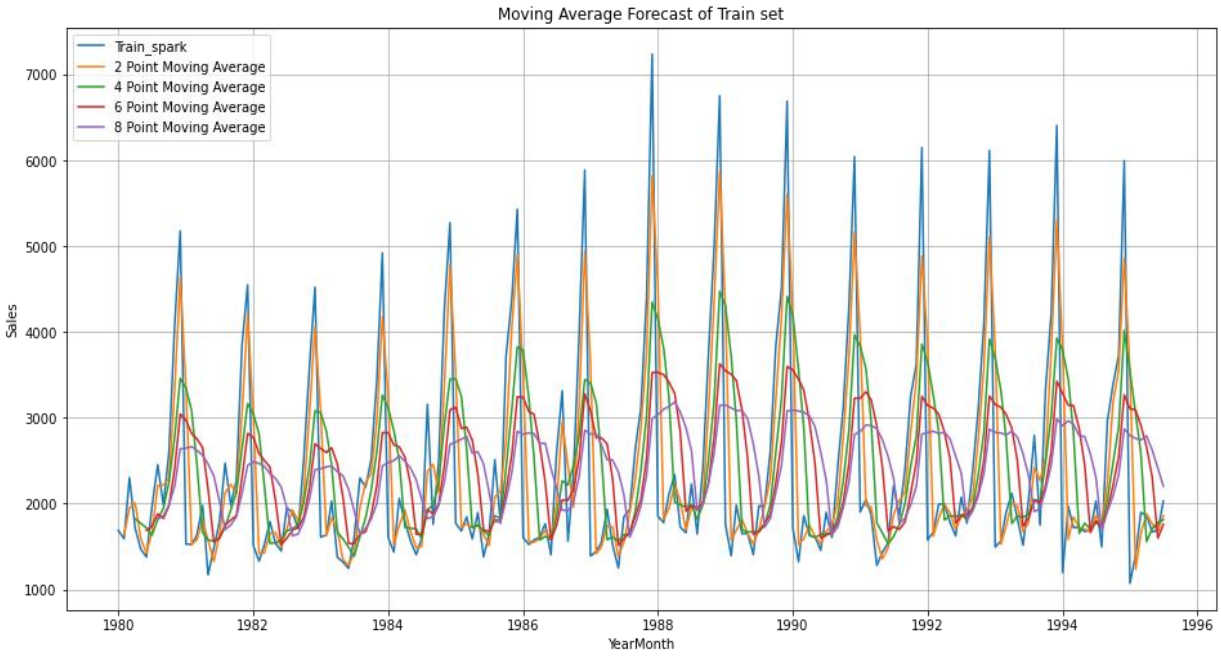

Fig. 07

- We can observe that 2 point Moving Average forecast curve is best fitting with the Train set curves.
- Let us check for the Test set also.
- Plot showing Train, Test, Moving average curves of Test set is as shown below:

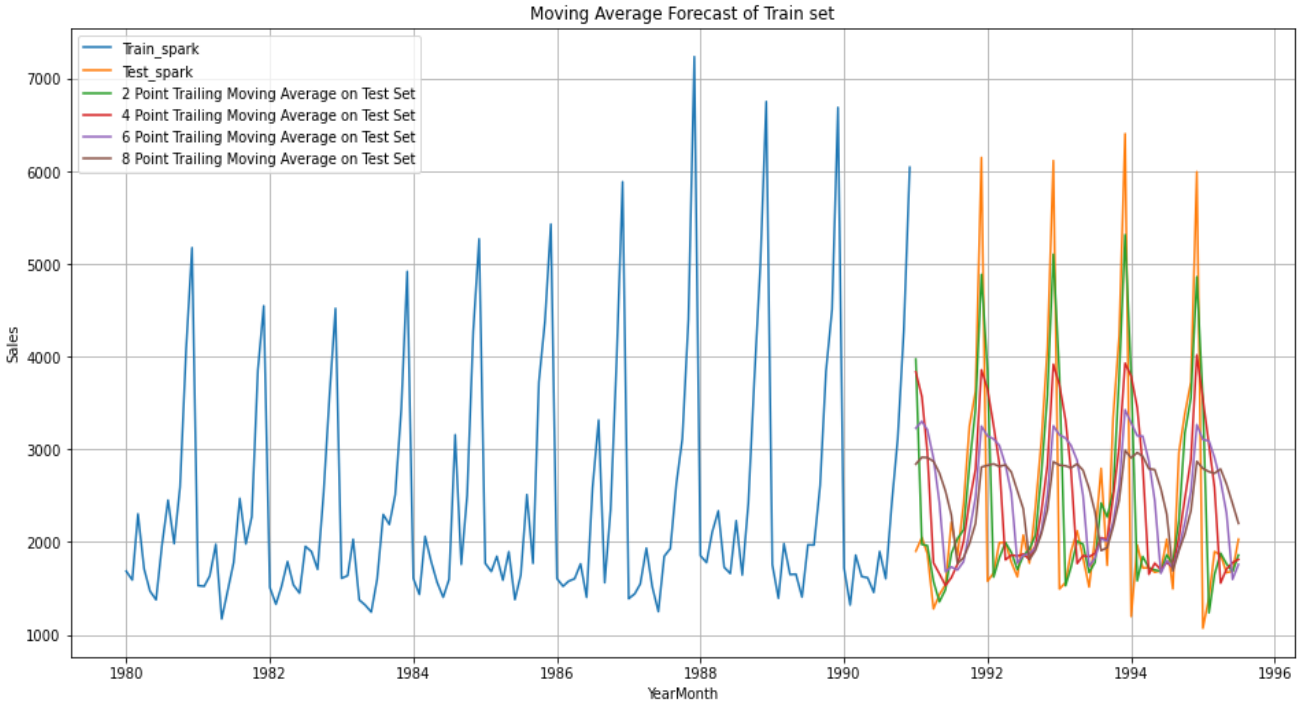

Fig. 08

- We can observe that 2 point Moving Average forecast curve is best fitting with the Train set curves.
- RMSE values for 2, 4, 6, 8 point Moving average curves are as shown below:

```
For 2 point Moving Average Model forecast on the spark Test Data,  RMSE is 813.401
For 2 point Moving Average Model forecast on the spark Test Data,  RMSE is 1156.590
For 2 point Moving Average Model forecast on the spark Test Data,  RMSE is 1283.927
For 2 point Moving Average Model forecast on the spark Test Data,  RMSE is 1342.568
```

Fig. 09

- 2 Point Moving average curve is the best fit among all.

5) Simple Exponential Smoothing:

- Let us fit the Simple Exponential Smoothing model on Training data set.
- Best parameters are as shown below:

```
{'smoothing_level': 0.07029459943040381,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1764.1004162520212,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 10

- Smoothing value is close 0, forecasts will farther from the actual values.

- Let us predict the Test set by using Simple Exponential Smoothing model.
- Sales value predicted by SES is ~2804
- Plot showing Train, Test, SES best fit line is as shown below:

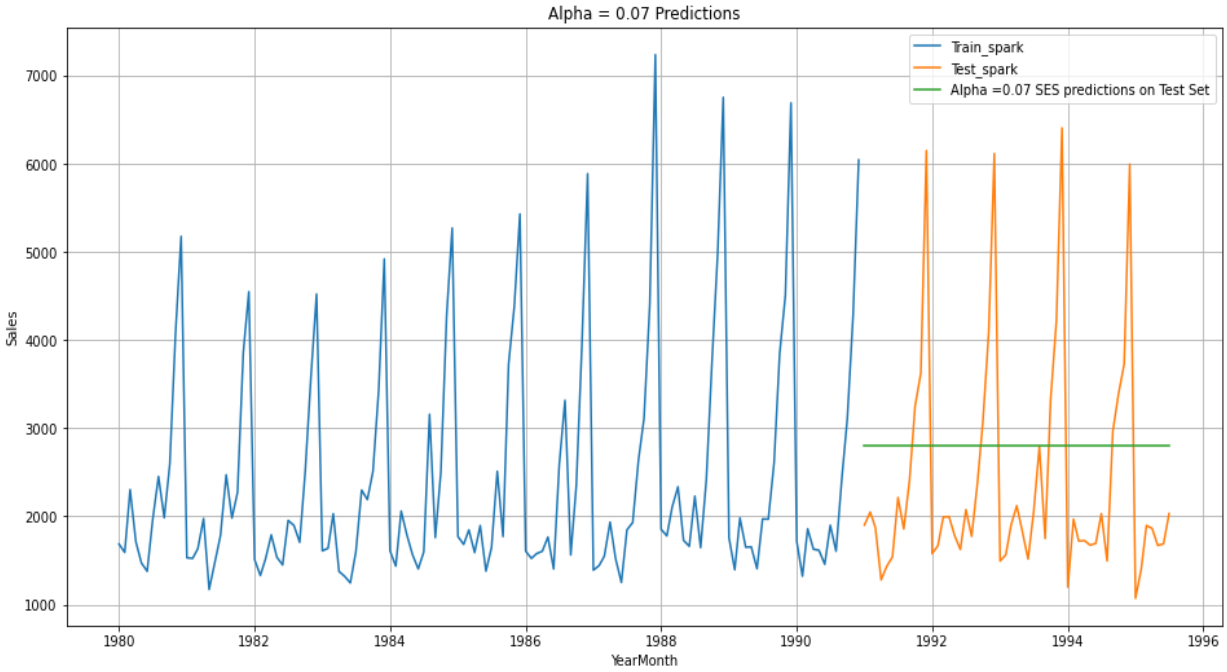

Fig. 11

- We can observe that SES model fit and Test set curves are not matching.
- RMSE value is ~1338.

6) Double Exponential Smoothing:

- Let us fit the Double Exponential Smoothing model on Training data set.
- Best parameters are as shown below:

```
{'smoothing_level': 0.6638769092832238,
 'smoothing_trend': 9.966251357628782e-05,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1502.5681711003654,
 'initial_trend': 29.020225552837097,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 12

- Smoothing value is 0.66, forecasts will moving closer to the actual values compared to SES Model.
- Let us predict the Test set by using Double Exponential Smoothing model. Sample of predicted values is as shown below:

```
1991-01-01    5330.501799
1991-02-01    5359.520204
1991-03-01    5388.538609
1991-04-01    5417.557013
1991-05-01    5446.575418
1991-06-01    5475.593823
1991-07-01    5504.612228
```

Fig. 13

- Plot showing Train, Test, SES, DES best fit line is as shown below:

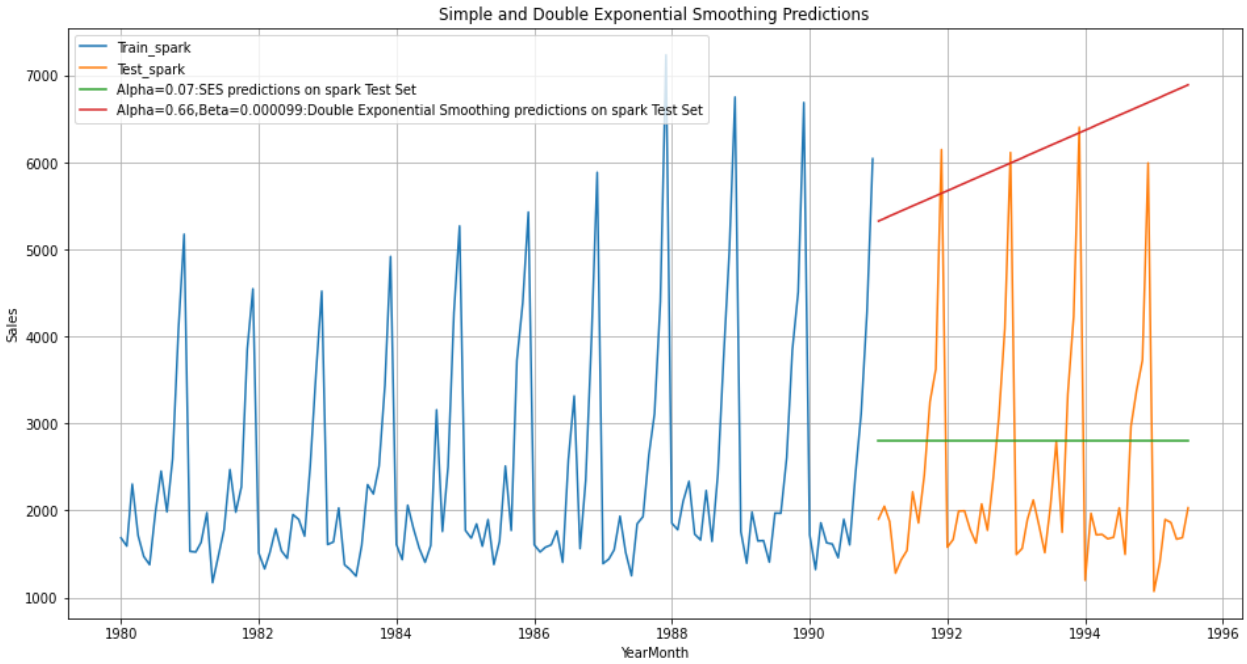

Fig. 14

- We can observe that DES model fit and Test set curves are not matching.
- RMSE value is ~3949.

7) Triple Exponential Smoothing:

- Let us fit the Triple Exponential Smoothing model on Training data set.
- Best parameters are as shown below:

```
{'smoothing_level': 0.10005373820823961,
 'smoothing_trend': 0.010034490652580457,
 'smoothing_seasonal': 0.5095957543425532,
 'damping_trend': nan,
 'initial_level': 2364.584774604334,
 'initial_trend': -0.016752880078245408,
 'initial_seasons': array([-653.82559323, -736.67734144, -368.25456128, -483.63906084,
        -826.15467946, -832.96819741, -386.3751117 ,   91.82676187,
        -261.32455153,  265.38968222, 1580.26233564, 2619.56221896]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 15

- Smoothing seasonal value is 0.50, forecasts will be moving closer to the actual values compared to SES Model.
- Let us predict the Test set by using Triple Exponential Smoothing model. Sample of predicted values is as shown below:

```
1991-01-01    1509.969093
1991-02-01    1205.343244
1991-03-01    1702.386113
1991-04-01    1548.514691
1991-05-01    1467.824074
1991-06-01    1287.109239
1991-07-01    1804.027662
```

Fig. 16

- Plot showing Train, Test, SES, DES, TES best fit line is as shown below:

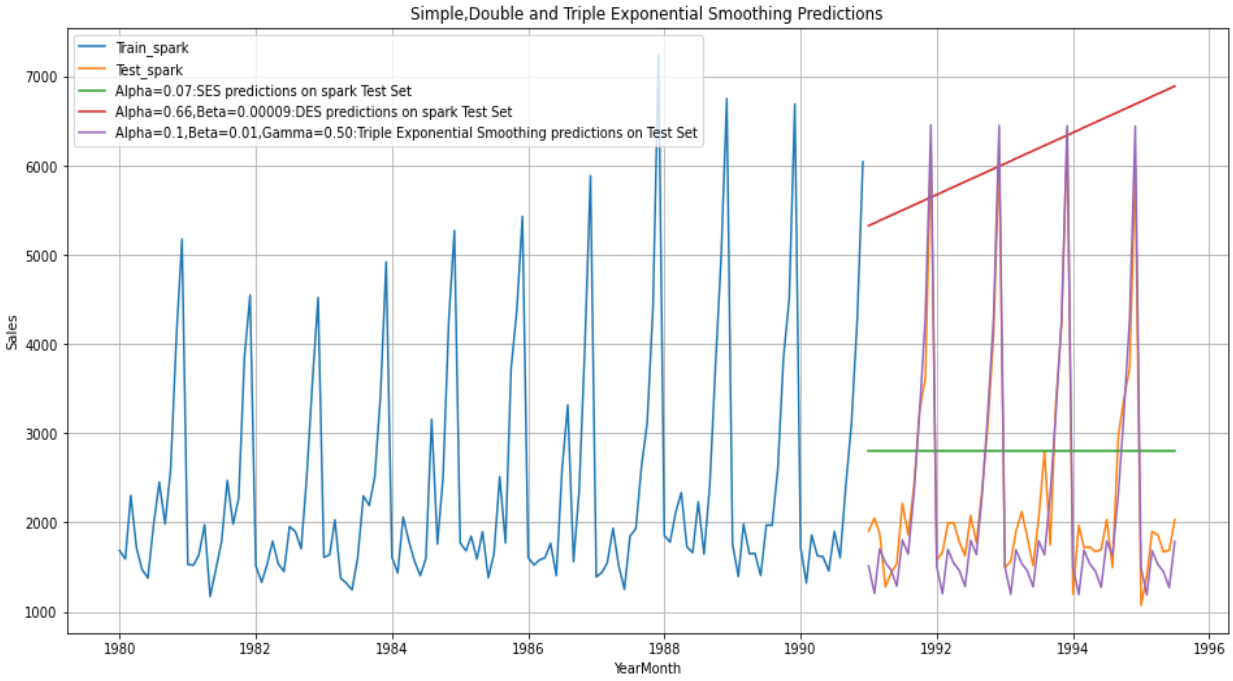

Fig. 17

- We can observe that TES model fit is mimicking the Test set in terms of trend and seasonality.
- RMSE value is ~379.

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Hypothesis testing for the Stationarity of the data:

Null hypothesis (H_0) - The Time Series has a unit root and is thus non-stationary.

Alternate hypothesis (H_1) - The Time Series does not have a unit root and is thus stationary.

- Let us perform ADF test on the dataset and check the results.

```
DF test statistic is -1.798
DF test p-value is 0.705595845993254
Number of lags used 12
```

Fig. 18

- p-value is greater than 0.05. We fail to reject the null hypothesis. So, at 5% significant level the Time Series is non-stationary.

Hypothesis testing for the Stationarity of the 1st order difference of the data:

Null hypothesis (H_0) - The 1st order difference of Time Series has a unit root and is thus non-stationary.
Alternate hypothesis (H_1) - The 1st order difference of Time Series does not have a unit root and is thus stationary.

- Now let us perform the ADF test on 1st order difference of the data set and check the results.

```
DF test statistic is -44.912
DF test p-value is 0.0
Number of lags used 10
```

Fig. 19

- p-value is less 0.05. We reject the null hypothesis. So, at 5% significant level the 1st order difference of the Time series is stationary.
- Plot of 1st order difference of the data is as shown below:

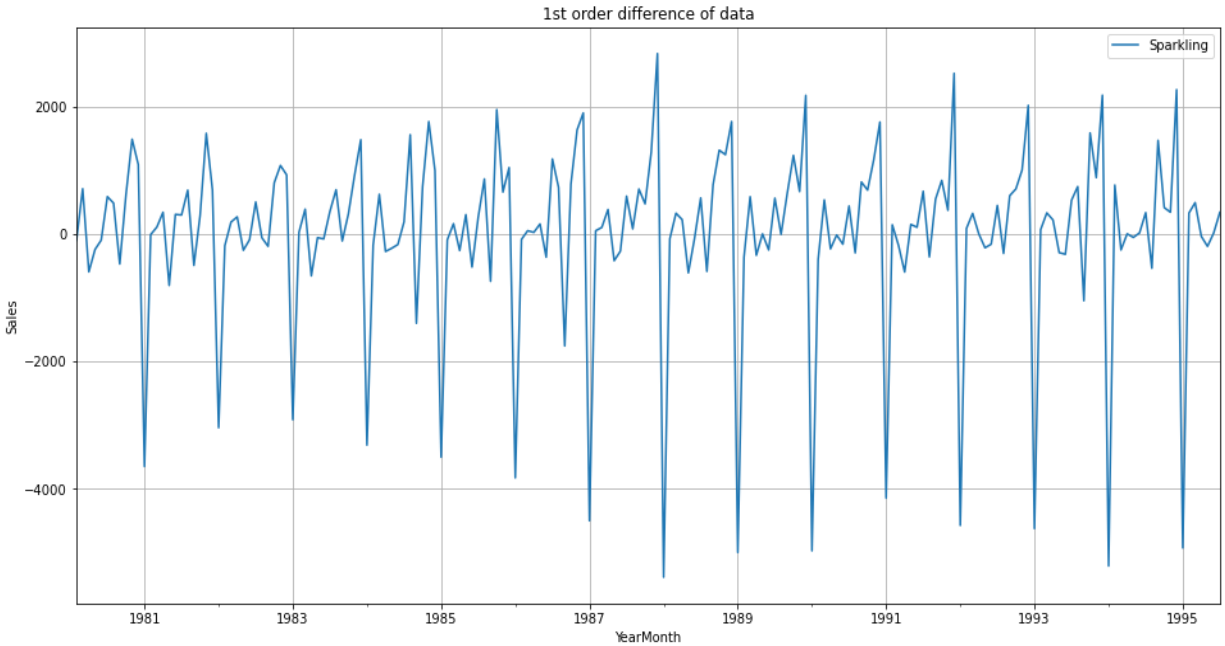

Fig. 20

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

ARIMA Model:

- Different combinations of parameters (p, d, q) defined for ARIMA model building are as shown below:

```
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Fig. 21

- Let us fit the Training data set with the defined p, d, q values and get the AIC scores. Sample of AIC scores obtained are as shown below:

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.060015591336
ARIMA(0, 1, 2) - AIC:2234.408323130674
```

Fig. 22

- AIC scores arranged in ascending order are shown below:

	param	AIC
10	(2, 1, 2)	2213.509213
15	(3, 1, 3)	2221.451354
14	(3, 1, 2)	2230.753752
11	(2, 1, 3)	2232.986167
9	(2, 1, 1)	2233.777626

Table. 12

- We can see that 2,1,2 combination of p, d, q is giving the lowest AIC score for the ARIMA model. Let us use this combination to fit the Training set and Test set. Results are shown below:

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1101.755
Date:	Sat, 22 Oct 2022	AIC	2213.509
Time:	18:56:23	BIC	2227.885
Sample:	01-01-1980 - 12-01-1990	HQIC	2219.351
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3121	0.046	28.781	0.000	1.223	1.401
ar.L2	-0.5593	0.072	-7.740	0.000	-0.701	-0.418
ma.L1	-1.9917	0.109	-18.214	0.000	-2.206	-1.777
ma.L2	0.9999	0.110	9.108	0.000	0.785	1.215
sigma2	1.099e+06	2e-07	5.51e+12	0.000	1.1e+06	1.1e+06

Ljung-Box (L1) (Q):	0.19	Jarque-Bera (JB):	14.46
Prob(Q):	0.67	Prob(JB):	0.00
Heteroskedasticity (H):	2.43	Skew:	0.61
Prob(H) (two-sided):	0.00	Kurtosis:	4.08

Fig. 23

- Plot diagnostics for the ARIMA model built are as shown below:

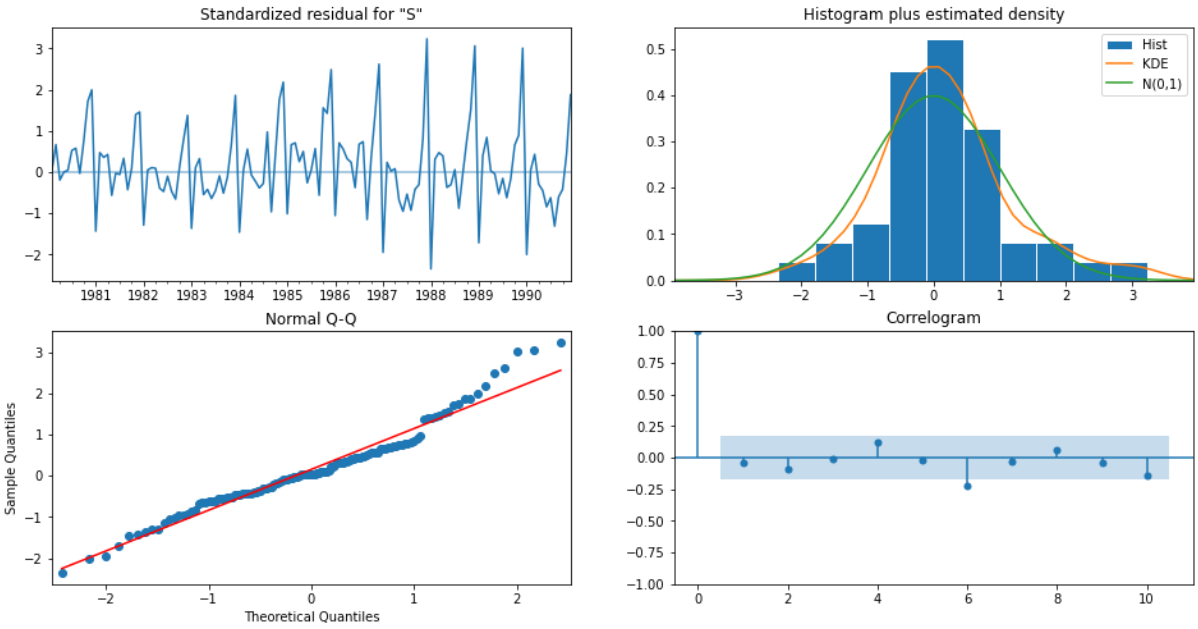

Fig. 24

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~1299.

SARIMA Model:

- Different combinations of parameters (p, d, q) & (P, D, Q, F) defined for ARIMA model building are as shown below:

```
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

Fig. 25

- Let us fit the Training data set with the defined (p, d, q) & (P, D, Q, F) values and get the AIC scores. Sample of AIC scores obtained are as shown below:

```
SARIMA(0, 1, 0)x(0, 0, 0, 6) - AIC:2251.3597196862966
SARIMA(0, 1, 0)x(0, 0, 1, 6) - AIC:2152.3780761716293
SARIMA(0, 1, 0)x(0, 0, 2, 6) - AIC:1955.6355536889566
```

Fig. 26

- AIC scores arranged in ascending order are shown below:

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	1629.051942
59	(0, 1, 3)	(2, 0, 3, 6)	1633.327869
123	(1, 1, 3)	(2, 0, 3, 6)	1633.988377
251	(3, 1, 3)	(2, 0, 3, 6)	1634.617459
63	(0, 1, 3)	(3, 0, 3, 6)	1635.054409

Table. 13

- We can see that (2,1,3) & (2,0,3,6) combination of (p, d, q) & (P, D, Q, F) is giving the lowest AIC score for the SARIMA model. Let us use this combination to fit the Training set and Test set. Results are shown below:

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 0, 3, 6)	Log Likelihood	-803.526
Date:	Mon, 17 Oct 2022	AIC	1629.052
Time:	22:05:39	BIC	1658.657
Sample:	01-01-1980 - 12-01-1990	HQIC	1641.058
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.7438	0.063	-27.736	0.000	-1.867	-1.621
ar.L2	-0.7863	0.068	-11.645	0.000	-0.919	-0.654
ma.L1	1.0841	0.290	3.732	0.000	0.515	1.653
ma.L2	-0.7528	0.123	-6.130	0.000	-0.994	-0.512
ma.L3	-0.8892	0.256	-3.468	0.001	-1.392	-0.387
ar.S.L6	-0.0117	0.029	-0.395	0.693	-0.069	0.046
ar.S.L12	1.0382	0.022	47.462	0.000	0.995	1.081
ma.S.L6	0.3731	0.261	1.430	0.153	-0.138	0.884
ma.S.L12	-0.7567	0.201	-3.760	0.000	-1.151	-0.362
ma.S.L18	0.1119	0.158	0.706	0.480	-0.199	0.422
sigma2	1.031e+05	4.97e-06	2.07e+10	0.000	1.03e+05	1.03e+05

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	14.99
Prob(Q):	0.93	Prob(JB):	0.00
Heteroskedasticity (H):	1.49	Skew:	0.38
Prob(H) (two-sided):	0.23	Kurtosis:	4.65

Fig. 27

- Plot diagnostics for the SARIMA model built are as shown below:

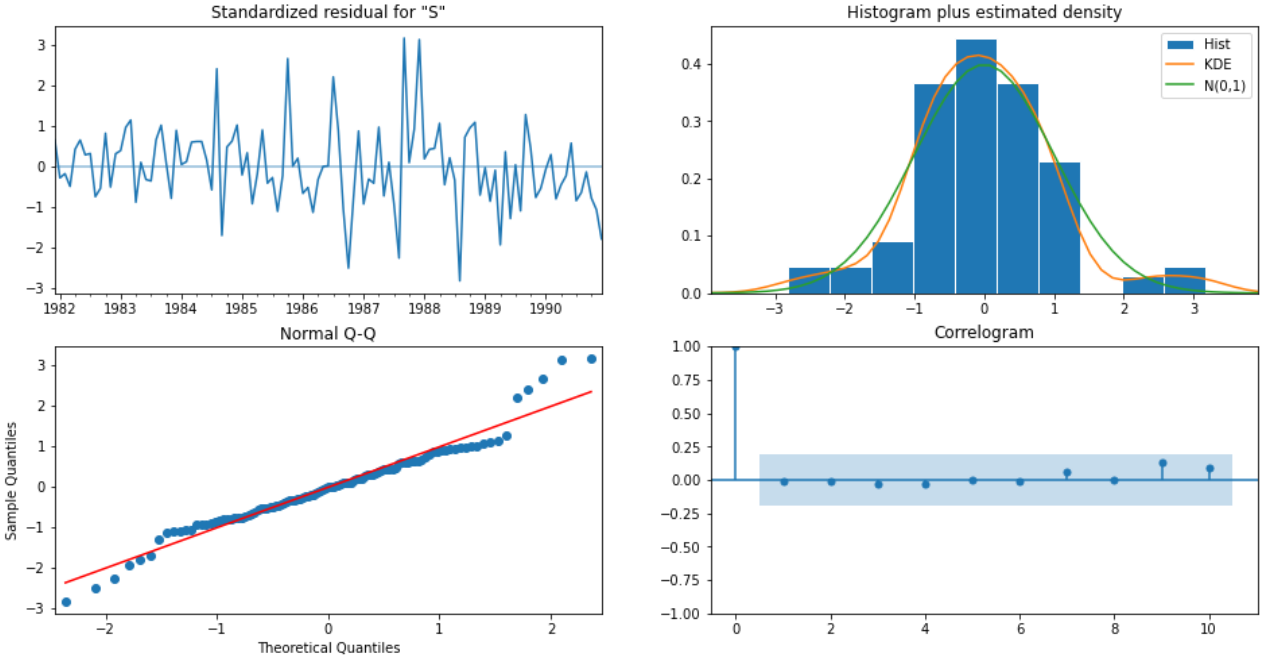

Fig. 28

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~826.

7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

ACF plot:

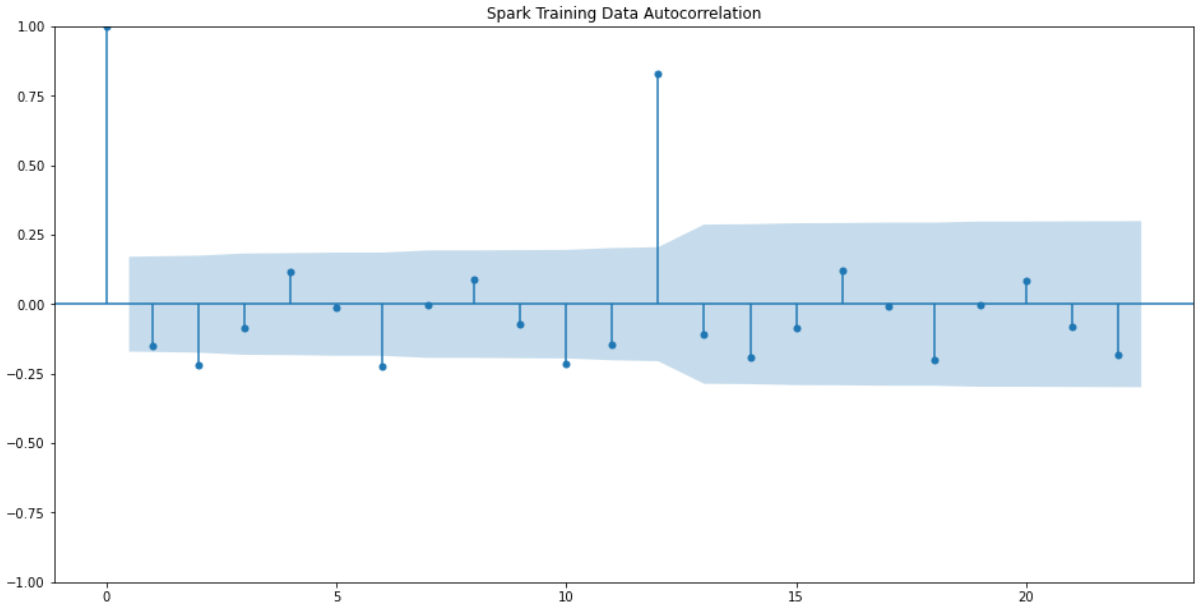

Fig. 29

- From above graph, q=2 and Q=4

PACF Plot:

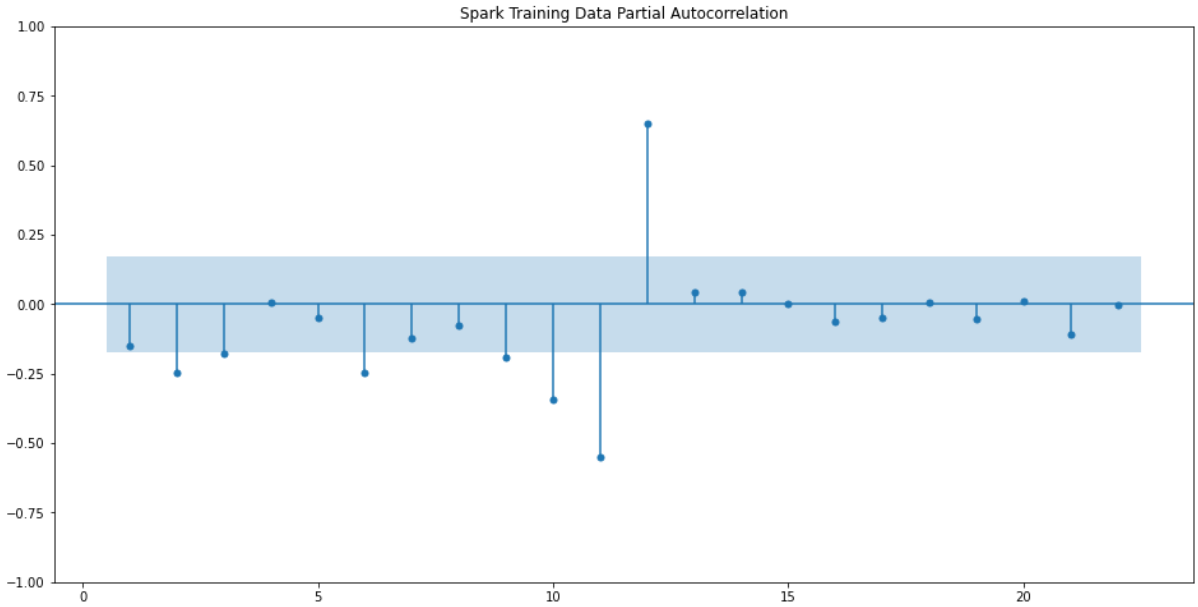

Fig. 30

- From graph, p=3 and P=0

- And we have already seen, 1st order difference of the data set is giving the stationary time series. So, d=1.

ARIMA Model:

- Let us fit the Training data set with the defined (p, d, q) = (3,1,2). Results are shown below:

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1109.377
Date:	Mon, 17 Oct 2022	AIC	2230.754
Time:	22:42:03	BIC	2248.005
Sample:	01-01-1980	HQIC	2237.764
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4330	0.042	-10.396	0.000	-0.515	-0.351
ar.L2	0.3259	0.112	2.903	0.004	0.106	0.546
ar.L3	-0.2411	0.071	-3.415	0.001	-0.379	-0.103
ma.L1	0.0194	0.127	0.152	0.879	-0.230	0.269
ma.L2	-0.9804	0.135	-7.243	0.000	-1.246	-0.715
sigma2	1.267e+06	1.94e-07	6.52e+12	0.000	1.27e+06	1.27e+06

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	4.60
Prob(Q):	0.89	Prob(JB):	0.10
Heteroskedasticity (H):	2.72	Skew:	0.37
Prob(H) (two-sided):	0.00	Kurtosis:	3.54

Fig. 31

- Plot diagnostics for the ARIMA model built are as shown below:

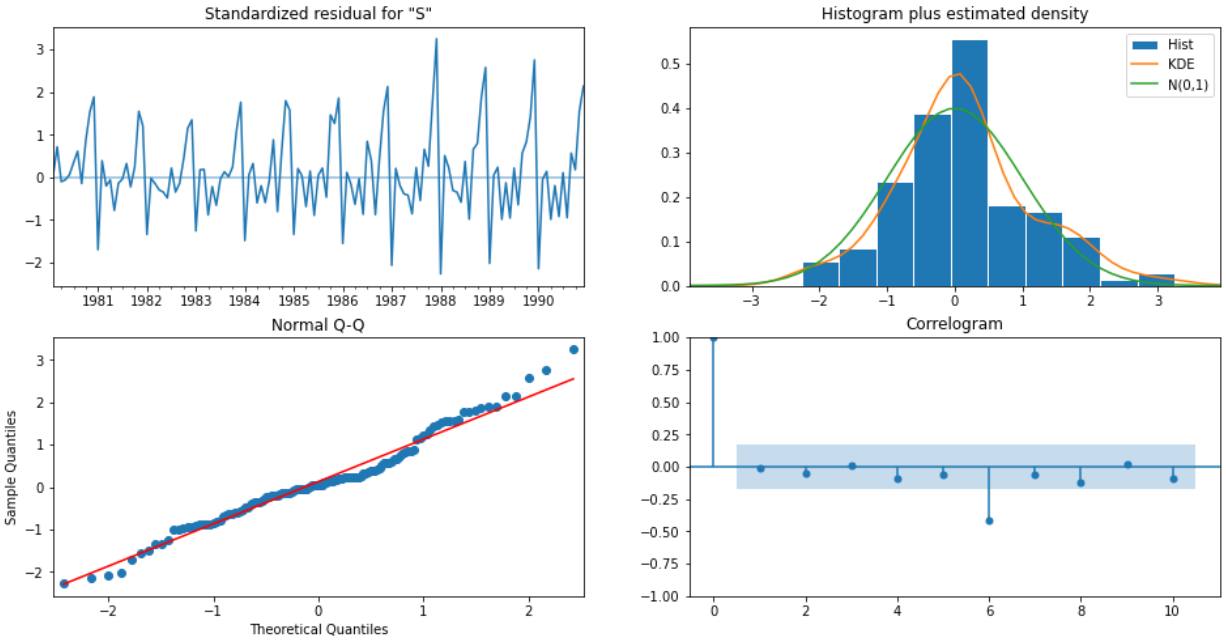

Fig. 32

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~1283.

SARIMA Model:

- Let us fit the Training data set with the defined (p, d, q) = (3,1,2) & (P, D, Q, F) = (0,1,4,6). Results are shown below:

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(0, 1, [1, 2, 3, 4], 6)	Log Likelihood	-771.591
Date:	Mon, 17 Oct 2022	AIC	1563.182
Time:	23:21:48	BIC	1589.032
Sample:	01-01-1980	HQIC	1573.638
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3301	0.112	11.832	0.000	1.110	1.550
ar.L2	-0.4617	0.197	-2.341	0.019	-0.848	-0.075
ar.L3	-0.2109	0.123	-1.712	0.087	-0.452	0.031
ma.L1	-1.8537	0.135	-13.755	0.000	-2.118	-1.590
ma.L2	1.0000	0.144	6.925	0.000	0.717	1.283
ma.S.L6	-1.5103	0.124	-12.194	0.000	-1.753	-1.268
ma.S.L12	1.3241	0.196	6.740	0.000	0.939	1.709
ma.S.L18	-0.8840	0.214	-4.130	0.000	-1.303	-0.464
ma.S.L24	0.3122	0.133	2.356	0.018	0.052	0.572
sigma2	3.855e+05	6.95e-07	5.55e+11	0.000	3.86e+05	3.86e+05

Ljung-Box (L1) (Q):	1.00	Jarque-Bera (JB):	11.27
Prob(Q):	0.32	Prob(JB):	0.00
Heteroskedasticity (H):	1.09	Skew:	0.61
Prob(H) (two-sided):	0.80	Kurtosis:	4.13

Fig. 33

- Plot diagnostics for the ARIMA model built are as shown below:

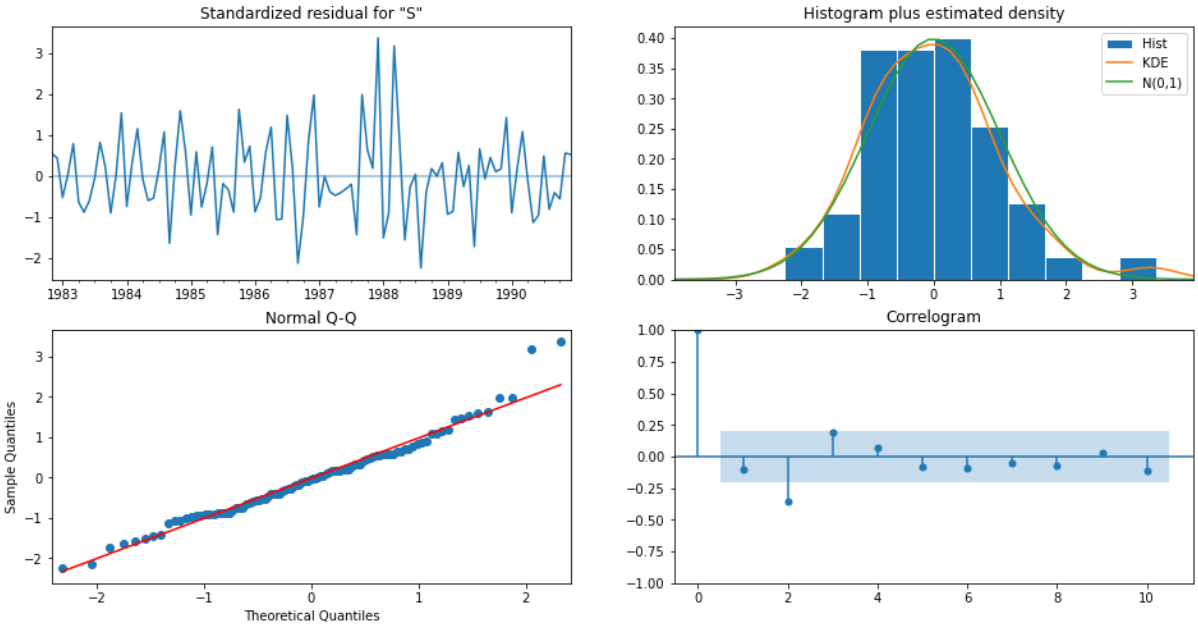

Fig. 34

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~1019.

8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Data frame with all the defined models and their test RMSE values:

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
8pointTrailingMovingAverage	1342.567772
SESModel	1338.012144
DESModel	3949.993290
TESModel	379.695686
ARIMA_pdq_Model	1299.979524
SARIMA_pdq_Model	826.659783
ARIMA_manual_Model	1283.352842
SARIMA_manual_Model	1019.873288

Table. 14

- TES model is best among all.

9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

- Let us predict the 12 months forecast by using the TES Model.
- Let us fit the whole data frame first and best parameters are as shown below:

```
{'smoothing_level': 0.07596713833847582,
 'smoothing_trend': 0.03256922042142542,
 'smoothing_seasonal': 0.37660763013263704,
 'damping_trend': nan,
 'initial_level': 2356.500976792558,
 'initial_trend': -1.0362742462267969,
 'initial_seasons': array([-636.25317961, -723.00028675, -398.67051497, -473.45456398,
       -808.43195611, -815.36867317, -384.24769271,   72.9999949 ,
       -237.46126013,  272.34548254, 1541.39087625, 2590.11216133]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 35

- Next 12 months forecast sales' values predicted by TES model are as shown below:

```
1995-08-01    1877.418973
1995-09-01    2405.272289
1995-10-01    3242.091582
1995-11-01    3922.174721
1995-12-01    6118.486885
1996-01-01    1262.602775
1996-02-01    1592.120997
1996-03-01    1831.635313
1996-04-01    1806.451718
1996-05-01    1651.704099
1996-06-01    1586.487882
1996-07-01    1976.989421
Freq: MS, dtype: float64
```

Fig. 36

- Plot showing whole data frame and 12 months forecast is as shown below:

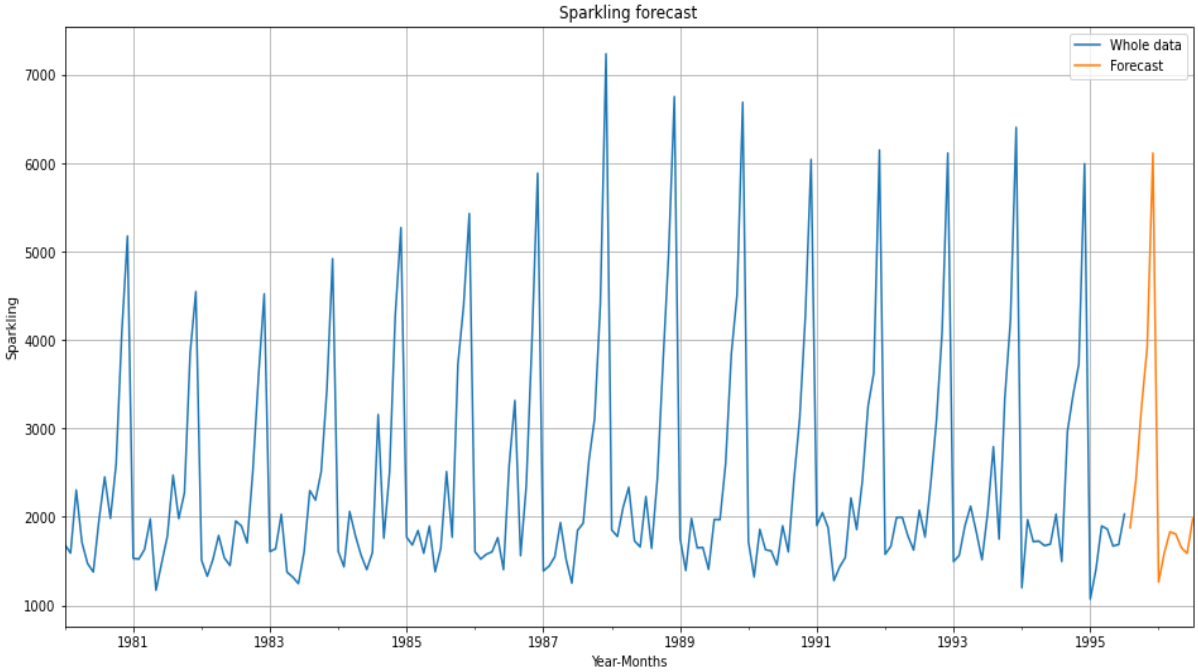

Fig. 37

10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- Forecast performed for 12 months is following trend that we observe in last 8 years.
- Dec month is having the highest sales of wines in all the years. Oct, Nov months have next highest sales. This is clearly showing wines sales are more in winter season.
- There is increase trend in sales up to 1988 year. After 1988, there is decrease trend in sales 1995. Forecast also shown same no. of sales for 1996 year also.
- By all above observations, we can say there is a demand for Sparkling wines because we do not observe rapid decreasing trend in wines' sales. There are approximately constant sales from 1991.
- So, company should continue their efforts maintain the consistency in sales.
- Company should also focus on increasing the sales over the years. It should try for increasing trend.

Rose Wines' Sales

1. Read the data as an appropriate Time Series data and plot the data.

Sample data frame as appropriate Time Series (TS) is as shown below:

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table. 15

- Data has been loaded as TS.
- Sales data is given Month-wise.

Plot of TS data:

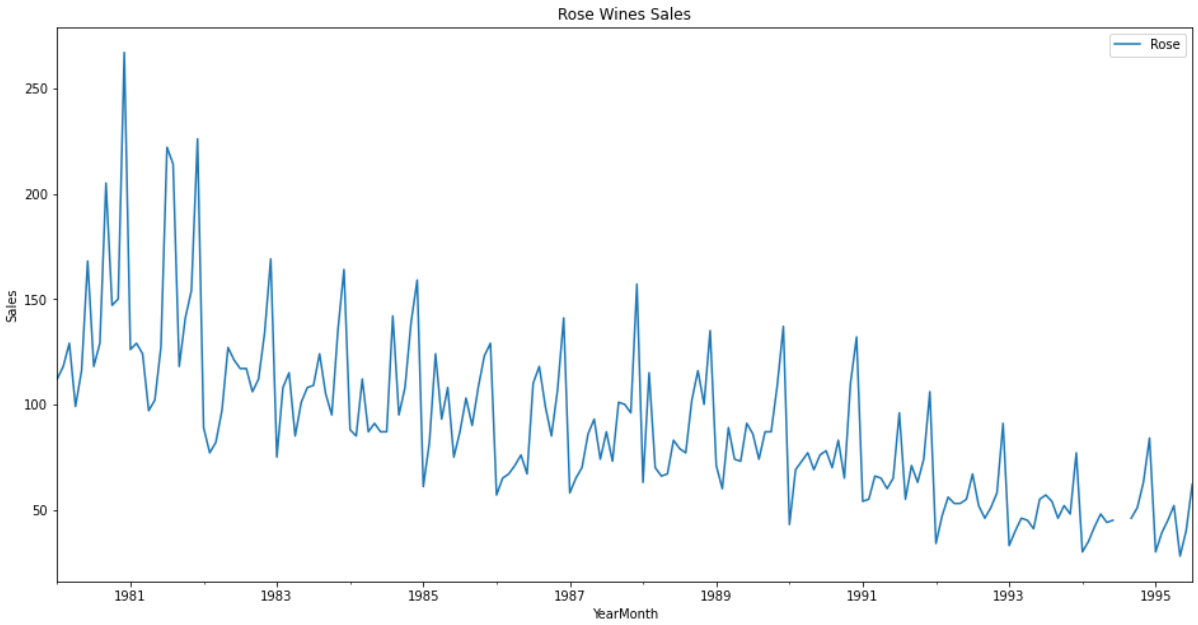

Fig. 38

Interpretation:

- There are decreasing trend in the TS data.
- There is no proper seasonality in the TS data.

2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

EDA:

First five records of TS data are as shown below:

	Rose
YearMonth	
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table. 16

Last five records of TS data are as shown below:

	Rose
YearMonth	
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

Table. 17

- Sales data is given from Jan-1980 to July-1995, i.e., total 187 months is given.
- Sales are not having proper seasonality by Month-wise sales year by year.

Data description: (for 187 months)

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

Table. 18

- Mean sales are ~90.
- Maximum sales are 267 (in May-1995).
- Minimum sales are 28 (in Dec-1980).
- There are 90 duplicated records as shown below.

YearMonth	Rose
1980-07-01	118.0
1980-08-01	129.0
1981-02-01	129.0
1981-09-01	118.0
1982-04-01	97.0
...	...
1994-11-01	63.0
1995-01-01	30.0
1995-03-01	45.0
1995-04-01	52.0
1995-06-01	40.0

Table. 19

- All of the duplicated records are significant.
- There are 2 null values for Jul-1994, Aug-1994 sales.
- Let us look at the plot of 1994 sales.

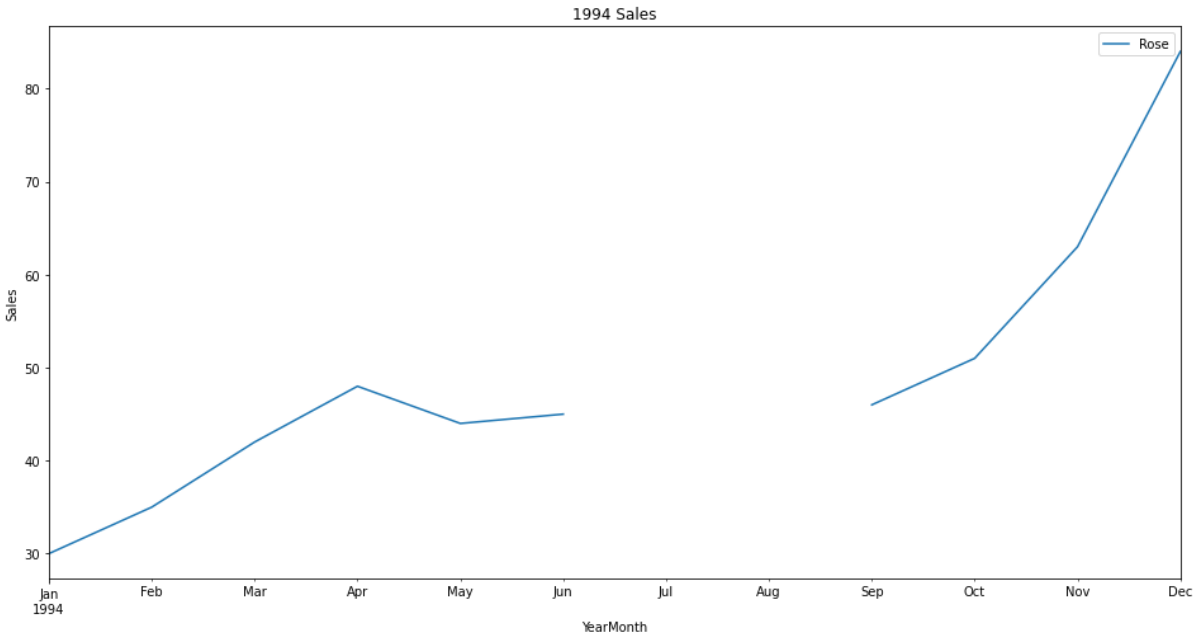

Fig. 39

- We can observe that there is slow increase trend. So, let us impute the Jul-1994, Aug-1994 sales by Jun-1994, Sep-1994 sales values respectively.
- Plot of 1994 wines' sales after imputing:

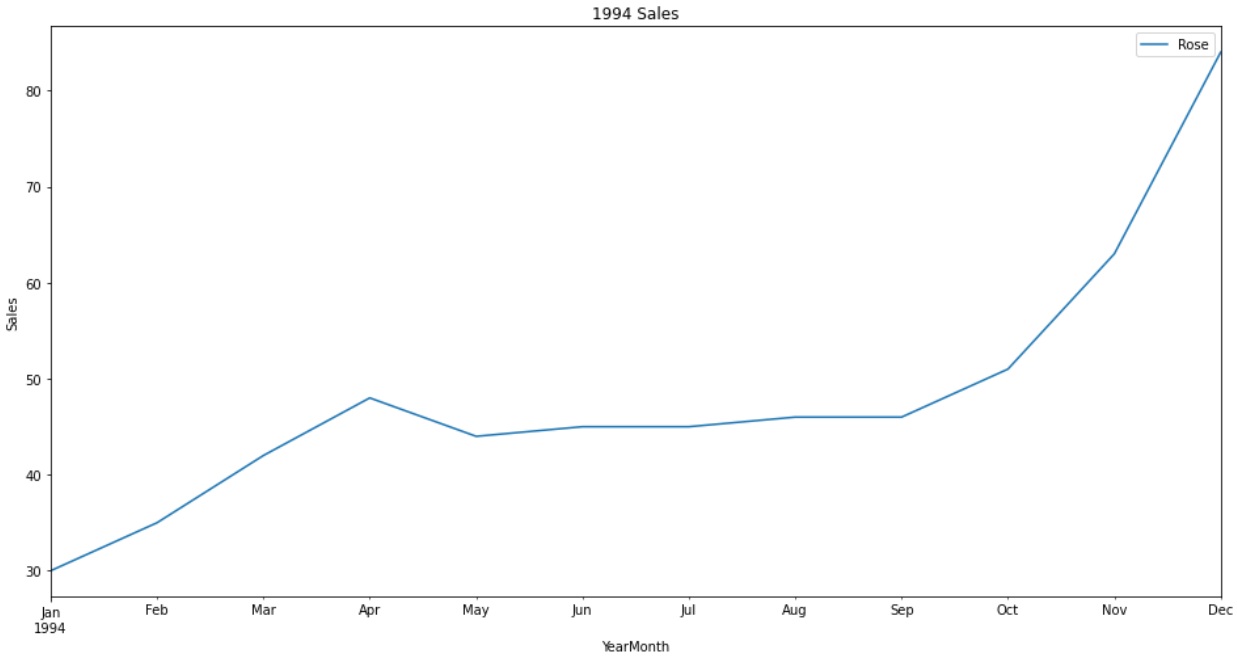

Fig. 40

- Null values treated successfully.

Decomposition:

- Let us perform decomposition of TS using additive method

Decomposition plot is as shown below:

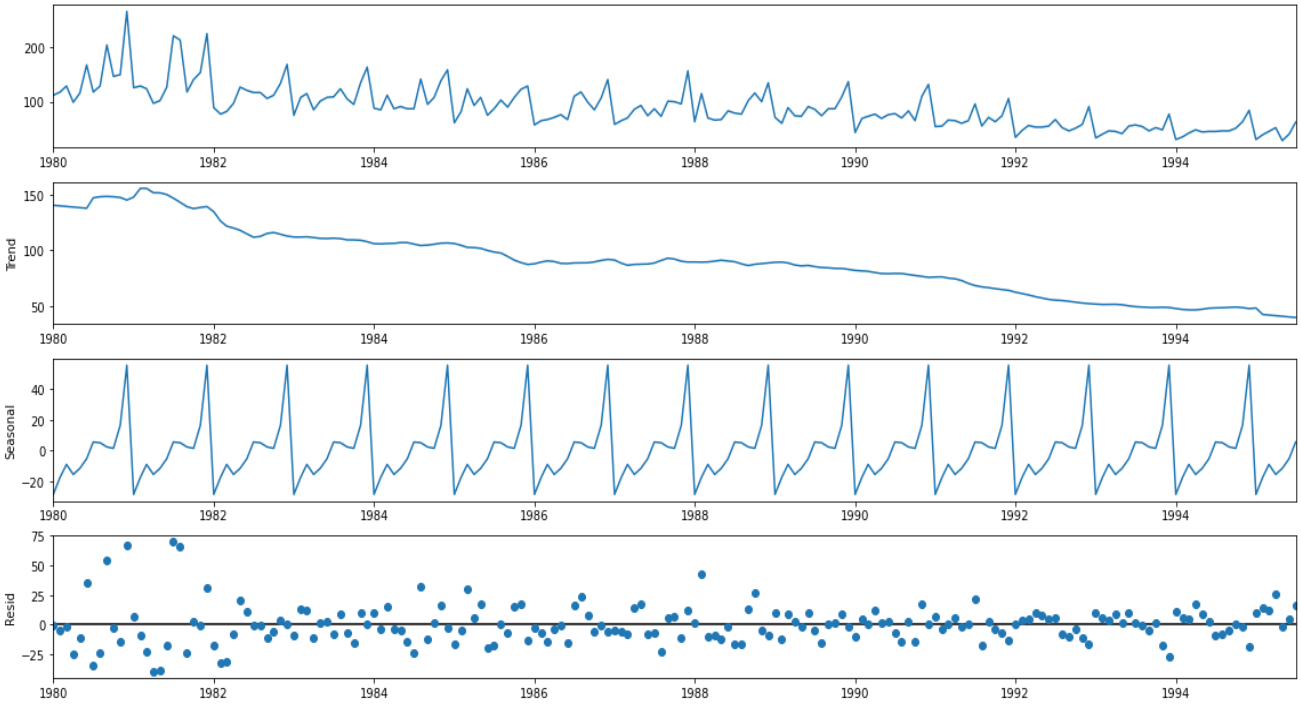

Fig. 41

Decomposition trend is as shown below:

```
YearMonth
1980-01-01    140.426734
1980-02-01    139.884824
1980-03-01    139.342914
1980-04-01    138.801004
1980-05-01    138.259094
                 ...
1995-03-01     41.799092
1995-04-01     41.257182
1995-05-01     40.715271
1995-06-01     40.173361
1995-07-01     39.631451
Name: trend, Length: 187, dtype: float64
```

Table. 20

Decomposition residuals are as shown below:

```
YearMonth
1980-01-01    -0.120661
1980-02-01    -4.692870
1980-03-01    -1.411407
1980-04-01   -24.413798
1980-05-01   -10.791189
                 ...
1995-03-01    12.132416
1995-04-01    26.130024
1995-05-01    -1.247367
1995-06-01     5.023680
1995-07-01    16.765542
Name: resid, Length: 187, dtype: float64
```

Table. 21

Decomposition seasonality is as shown below:

```
YearMonth
1980-01-01   -28.306074
1980-02-01   -17.191954
1980-03-01    -8.931507
1980-04-01   -15.387206
1980-05-01   -11.467905
                 ...
1995-03-01    -8.931507
1995-04-01   -15.387206
1995-05-01   -11.467905
1995-06-01    -5.197041
1995-07-01     5.603007
Name: seasonal, Length: 187, dtype: float64
```

Table. 22

Evaluating the TS Decomposition: (for July-1995)

- Sales by TS decomposition are 61.90.
- Actual sales are 62.
- TS decomposition of Sparkling wines sales by additive method working well.

3. Split the data into training and test. The test data should start in 1991.

Data split:

- Train set has 132 records and Test set has 55 records i.e., in ~70:30 ratio.
- Sample Train data is as shown below:

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table. 23

- Sample Test data is as shown below:

YearMonth	Rose
1991-01-01	54.0
1991-02-01	55.0
1991-03-01	66.0
1991-04-01	65.0
1991-05-01	60.0

Table. 24

- We can see that Test data set is taken from Jan-1991.

4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other additional models such as regression, naïve forecast models, simple average models, moving average models should also be built on the training data and check the performance on the test data using RMSE.

1) Linear Regression Model:

- Let us fit the Linear Regression Model on Training data set.

```
▾ LinearRegression
LinearRegression()
```

Fig. 42

- Plot showing Train, Test, Linear Regression Model best fit line is as shown below:

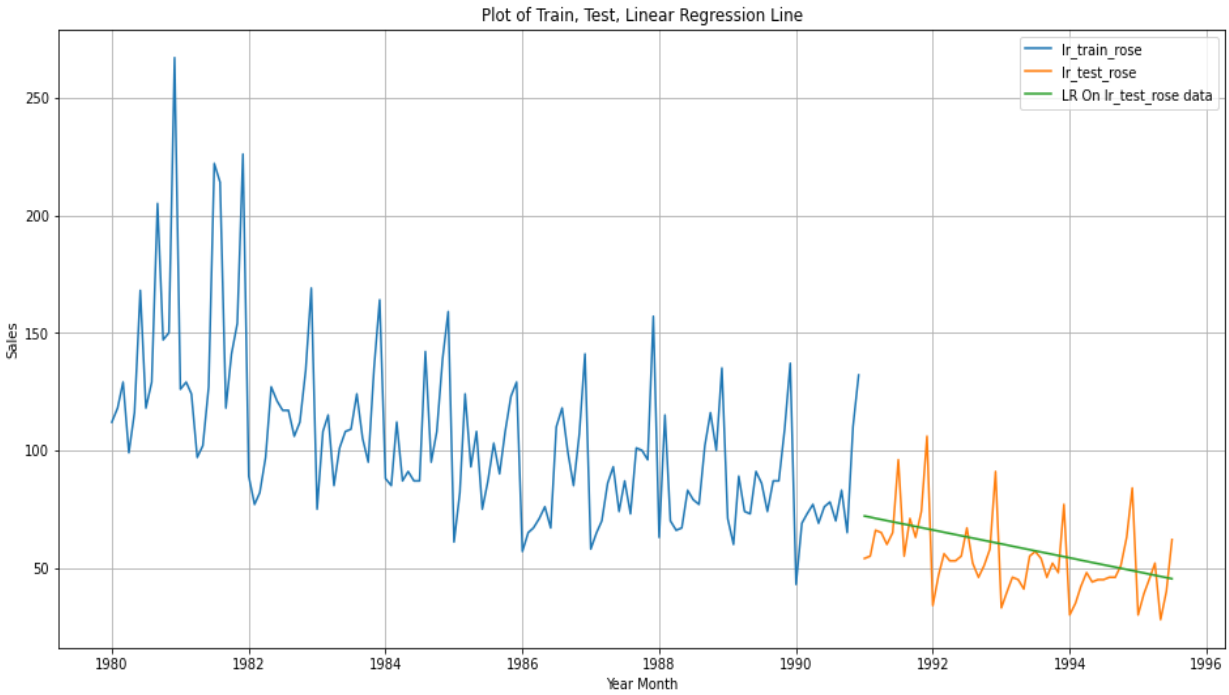

Fig. 43

- We can observe that Linear Regression model line and Test set curves are not matching.
- RMSE value is ~15.

2) Naïve Approach:

- Let us use the last value of the Training data set to predict the test set.
- Last value of the Training data set 132.
- Plot showing Train, Test, Naïve forecast line is as shown below:

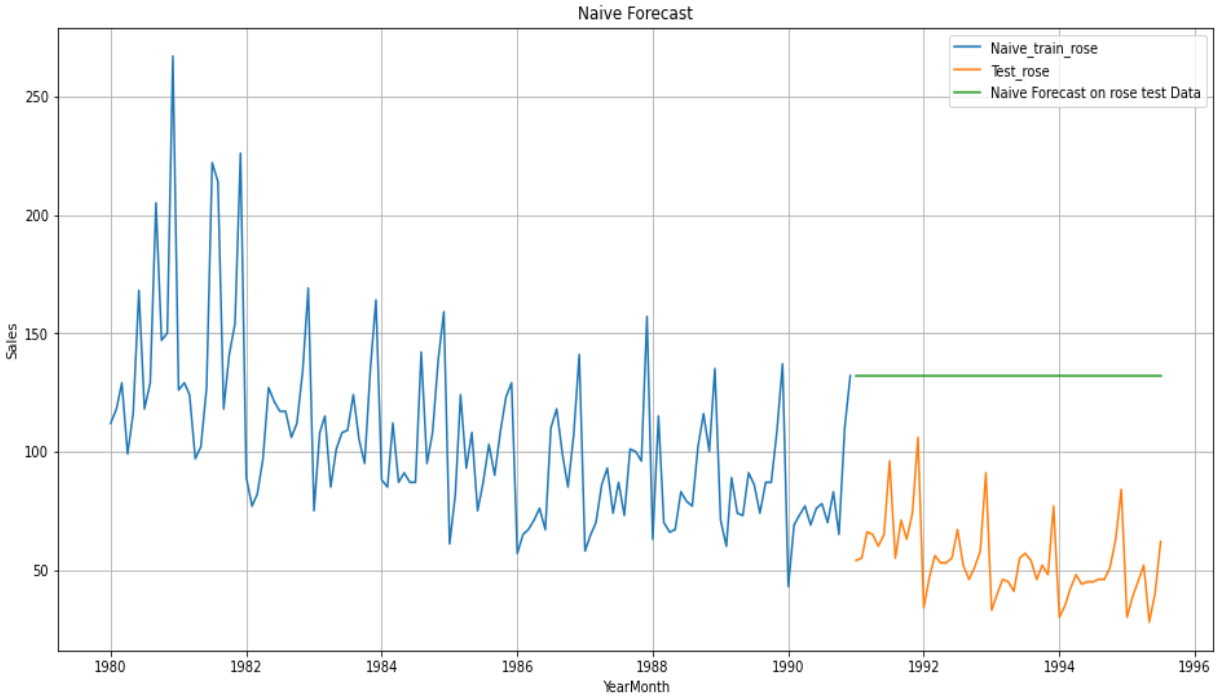

Fig. 44

- We can observe that Naïve forecast line and Test set curves are not matching.
- RMSE value is ~79.

3) Simple Average Model:

- Let us use the Average sales' value of the Test data set to predict the test set.
- Average sales' value of the Test data set is ~105.
- Plot showing Train, Test, Simple average line is as shown below:

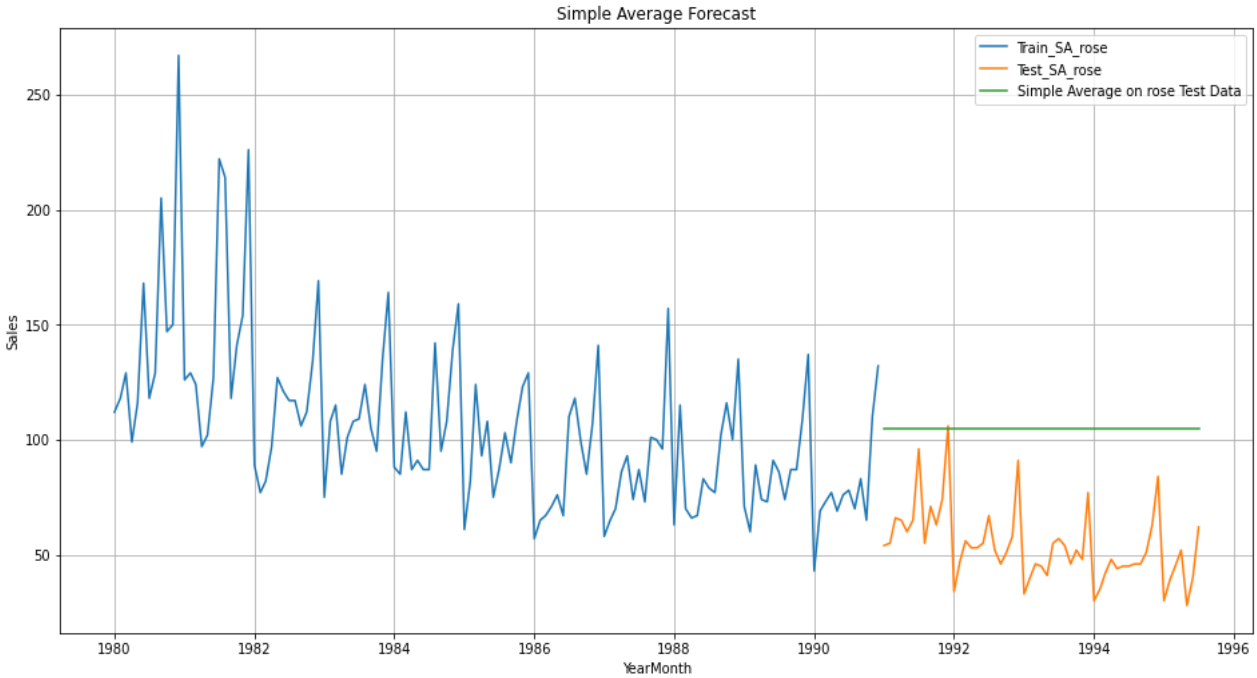

Fig. 45

- We can observe that Simple average forecast line and Test set curves are not matching.
- RMSE value is ~53.

4) Moving Average Model:

- Let us use the 2, 4, 6, 8 Rolling mean sales' values of the Train data set to predict the Test set.
- Sample data frame showing 2, 4, 6, 8 Moving Average sales' values of the Training data set is shown below:

YearMonth	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_8
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN

Table. 25

- Plot showing Train, Test, Moving average curves of Train set is as shown below:

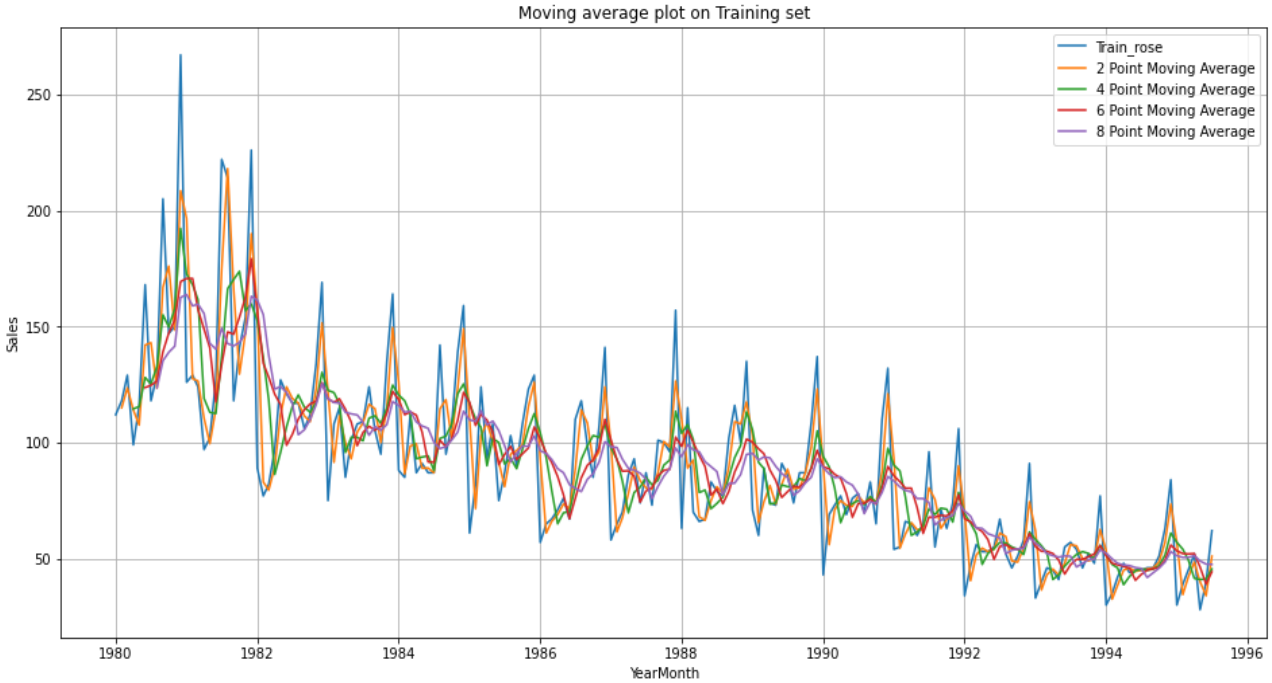

Fig. 46

- We can observe that 2 point Moving Average forecast curve is best fitting with the Train set curves.
- Let us check for the Test set also.
- Plot showing Train, Test, Moving average curves of Test set is as shown below:

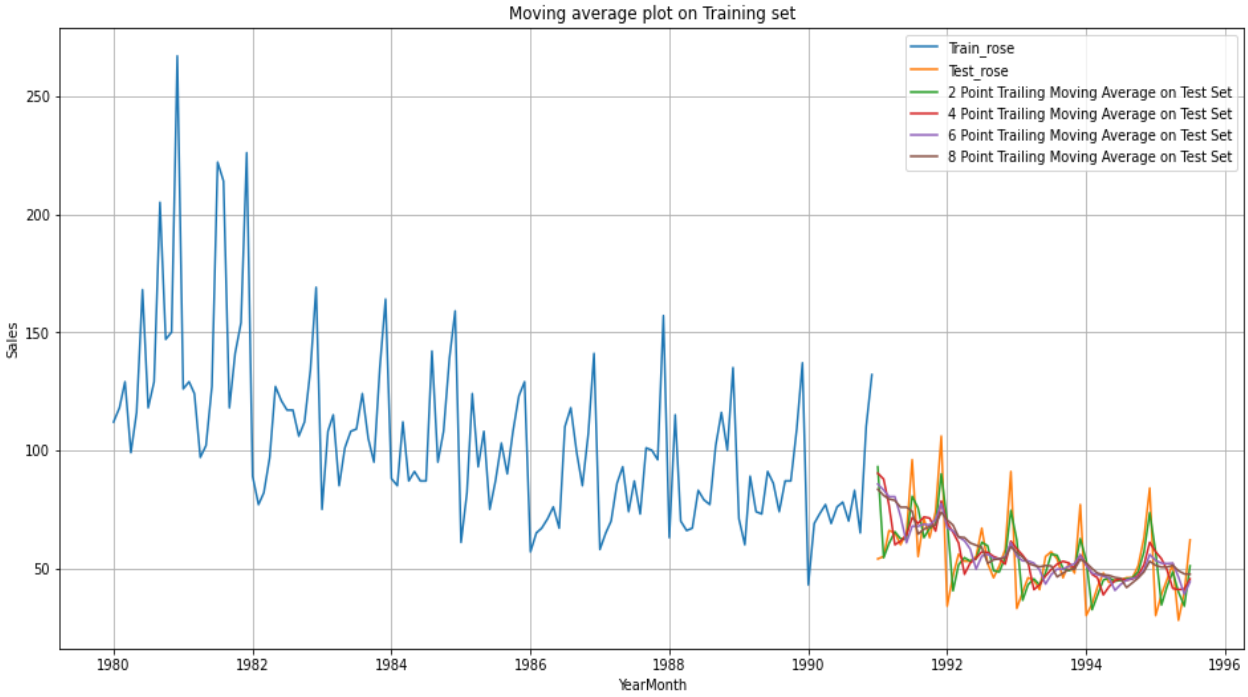

Fig. 47

- We can observe that 2 point Moving Average forecast curve is best fitting with the Train set curves.
- RMSE values for 2, 4, 6, 8 point Moving average curves are as shown below:

```
For 2 point Moving Average Model forecast on the rose Test Data,  RMSE is 11.529
For 4 point Moving Average Model forecast on the rose Test Data,  RMSE is 14.451
For 6 point Moving Average Model forecast on the rose Test Data,  RMSE is 14.568
For 8 point Moving Average Model forecast on the rose Test Data,  RMSE is 14.807
```

Fig. 48

- 2 Point Moving average curve is the best fit among all.

5) Simple Exponential Smoothing:

- Let us fit the Simple Exponential Smoothing model on Training data set.
- Best parameters are as shown below:

```
{'smoothing_level': 0.09874933517484011,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 134.38703609891138,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 49

- Smoothing value is close 0, forecasts will be farther from the actual values.

- Let us predict the Test set by using Simple Exponential Smoothing model.
- Sales value predicted by SES is ~87.
- Plot showing Train, Test, SES best fit line is as shown below:

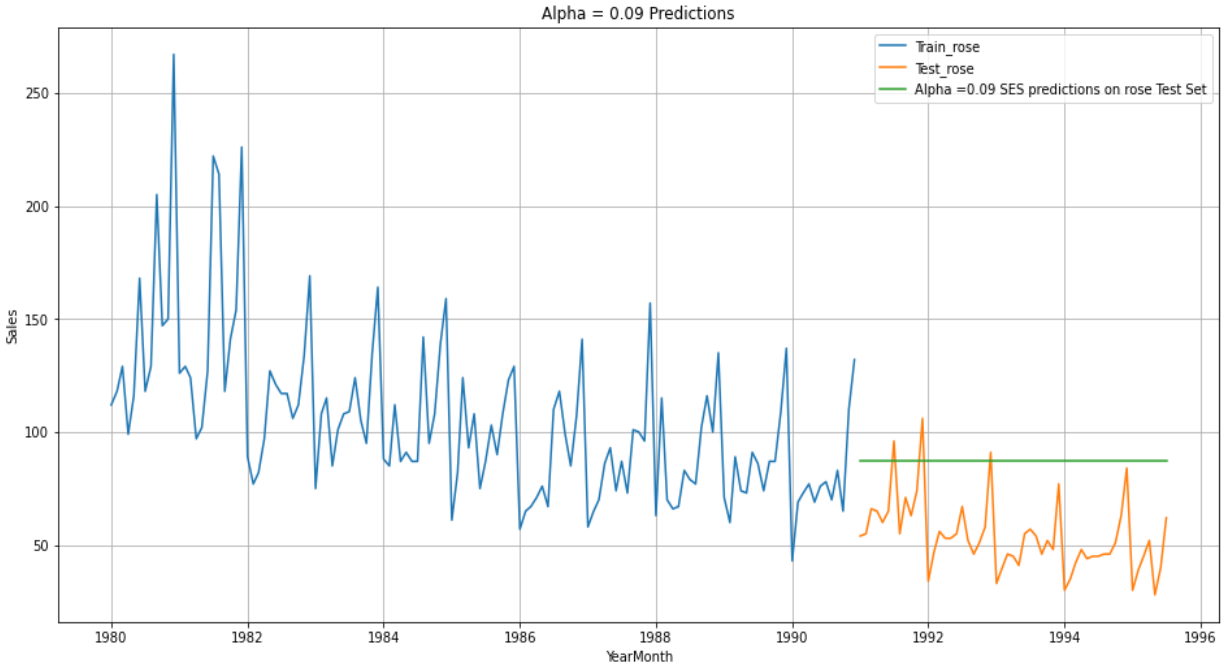

Fig. 50

- We can observe that SES model fit and Test set curves are not matching.
- RMSE value is ~36.

6) Double Exponential Smoothing:

- Let us fit the Double Exponential Smoothing model on Training data set.
- Best parameters are as shown below:

```
{'smoothing_level': 1.9086427682180844e-08,
 'smoothing_trend': 7.302464353829351e-09,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 137.81629861505857,
 'initial_trend': -0.4943753249082896,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 51

- Smoothing value is close to 0, forecasts will be farther from the actual values.
- Let us predict the Test set by using Double Exponential Smoothing model. Sample of predicted values is as shown below:

1991-01-01	72.064380		
1991-02-01	71.570005	1991-01-01	5330.501799
1991-03-01	71.075630	1991-02-01	5359.520204
1991-04-01	70.581254	1991-03-01	5388.538609
1991-05-01	70.086879	1991-04-01	5417.557013
1991-06-01	69.592504	1991-05-01	5446.575418
1991-07-01	69.098128	1991-06-01	5475.593823
1991-08-01	68.603753	1991-07-01	5504.612228

Fig. 52

- Plot showing Train, Test, SES, DES best fit line is as shown below:

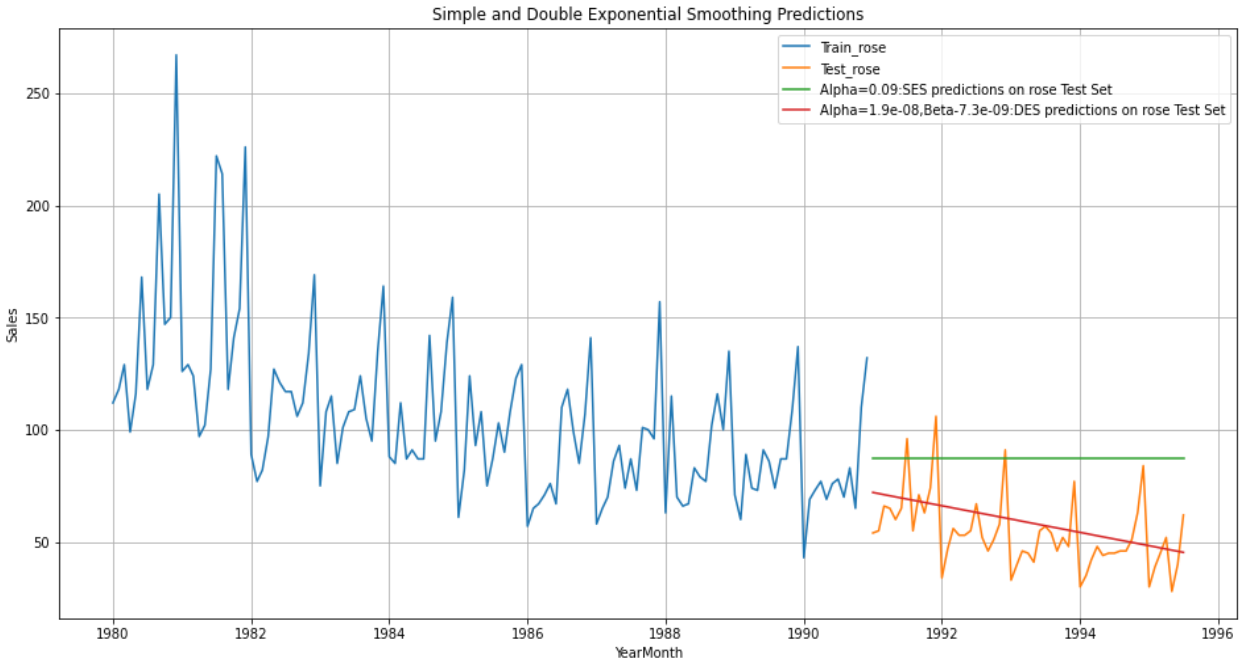

Fig. 53

- We can observe that DES model fit and Test set curves are not matching.
- RMSE value is ~15.

7) Triple Exponential Smoothing:

- Let us fit the Triple Exponential Smoothing model on Training data set.
- Best parameters are as shown below:

```
{'smoothing_level': 0.08830330642635406,
 'smoothing_trend': 6.730635331927582e-05,
 'smoothing_seasonal': 0.004455138229351625,
 'damping_trend': nan,
 'initial_level': 146.88752868155674,
 'initial_trend': -0.5492163940406024,
 'initial_seasons': array([-31.12207537, -18.81171138, -10.86052241, -21.52235816,
       -12.68359535,  -7.17529564,   2.7456236 ,   8.84900094,
         4.85724354,   2.9520333 ,  21.05004912,  63.29916317]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 54

- Let us predict the Test set by using Triple Exponential Smoothing model. Sample of predicted values is as shown below:

1991-01-01	42.672382
1991-02-01	54.439917
1991-03-01	61.841877
1991-04-01	50.636896
1991-05-01	58.918913
1991-06-01	63.870294
1991-07-01	73.240626

Fig. 55

- Plot showing Train, Test, SES, DES, TES best fit line is as shown below:

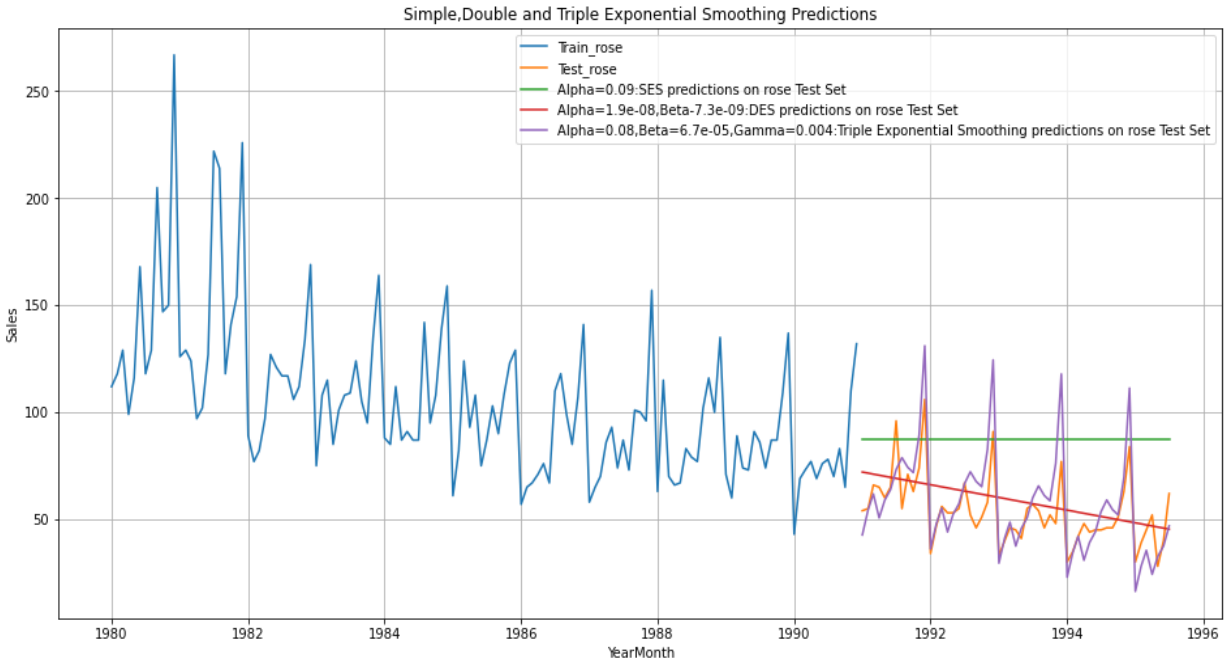

Fig. 56

- We can observe that TES model fit is mimicking the Test set in terms of trend and seasonality.
- RMSE value is ~14.

5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Hypothesis testing for the Stationarity of the data:

Null hypothesis (H_0) - The Time Series has a unit root and is thus non-stationary.
Alternate hypothesis (H_1) - The Time Series does not have a unit root and is thus stationary.

- Let us perform ADF test on the dataset and check the results.

```
DF test statistic is -2.240
DF test p-value is 0.4675416978196642
Number of lags used 13
```

Fig. 57

- p-value is greater than 0.05. We fail to reject the null hypothesis. So, at 5% significant level the Time Series is non-stationary.

Hypothesis testing for the Stationarity of the 1st order difference of the data:

Null hypothesis (H_0) - The 1st order difference of Time Series has a unit root and is thus non-stationary.
Alternate hypothesis (H_1) - The 1st order difference of Time Series does not have a unit root and is thus stationary.

- Now let us perform the ADF test on 1st order difference of the data set and check the results.

```
DF test statistic is -8.162
DF test p-value is 3.0125343211338084e-11
Number of lags used 12
```

Fig. 58

- p-value is less 0.05. We reject the null hypothesis. So, at 5% significant level the 1st order difference of the Time series is stationary.
- Plot of 1st order difference of the data is as shown below:

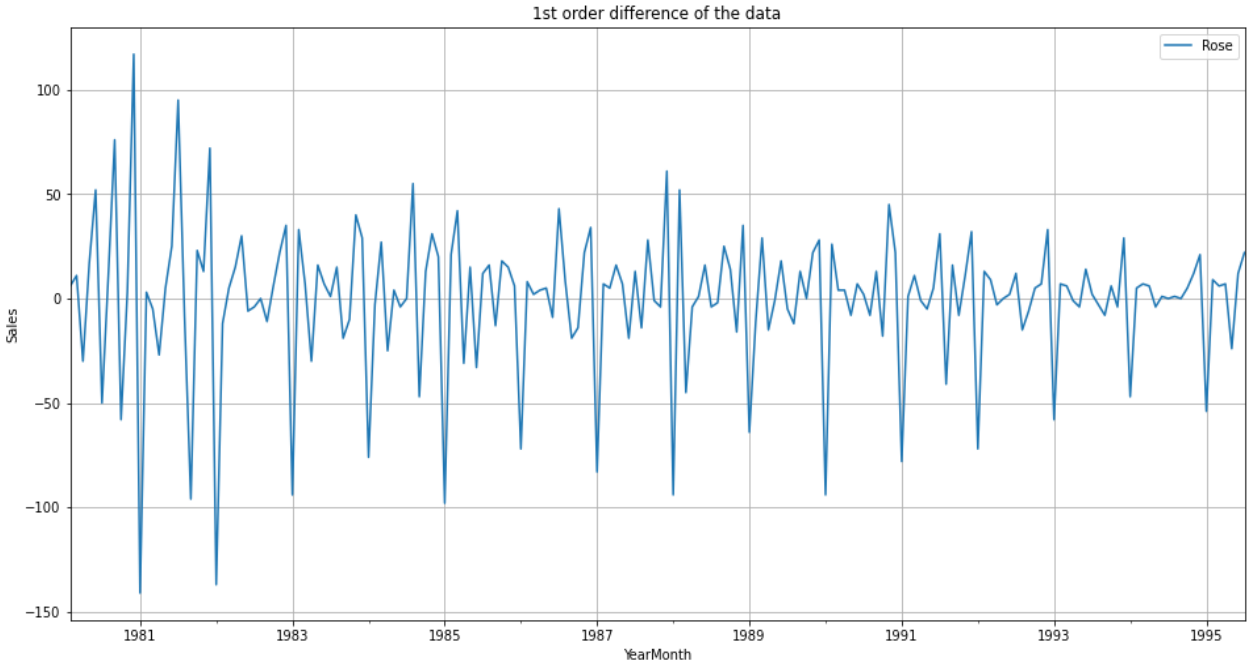

Fig. 59

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

ARIMA Model:

- Different combinations of parameters (p, d, q) defined for ARIMA model building are as shown below:

```
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Fig. 60

- Let us fit the Training data set with the defined p, d, q values and get the AIC scores. Sample of AIC scores obtained are as shown below:

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.3098319748315
ARIMA(0, 1, 2) - AIC:1279.6715288535743
```

Fig. 61

- AIC scores arranged in ascending order are shown below:

	param	AIC
11	(2, 1, 3)	1274.695136
15	(3, 1, 3)	1278.661305
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376

Table. 26

- We can see that 2,1,3 combination of p, d, q is giving the lowest AIC score for the ARIMA model. Let us use this combination to fit the Training set and Test set. Results are shown below:

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(2, 1, 3)	Log Likelihood	-631.348
Date:	Mon, 17 Oct 2022	AIC	1274.695
Time:	22:11:32	BIC	1291.946
Sample:	01-01-1980 - 12-01-1990	HQIC	1281.705
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.6779	0.084	-20.050	0.000	-1.842	-1.514
ar.L2	-0.7289	0.084	-8.711	0.000	-0.893	-0.565
ma.L1	1.0448	0.662	1.578	0.114	-0.253	2.342
ma.L2	-0.7718	0.135	-5.719	0.000	-1.036	-0.507
ma.L3	-0.9047	0.601	-1.506	0.132	-2.082	0.273
sigma2	858.1545	557.099	1.540	0.123	-233.740	1950.049

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	24.45
Prob(Q):	0.88	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.71
Prob(H) (two-sided):	0.00	Kurtosis:	4.57

Fig. 62

- Plot diagnostics for the ARIMA model built are as shown below:

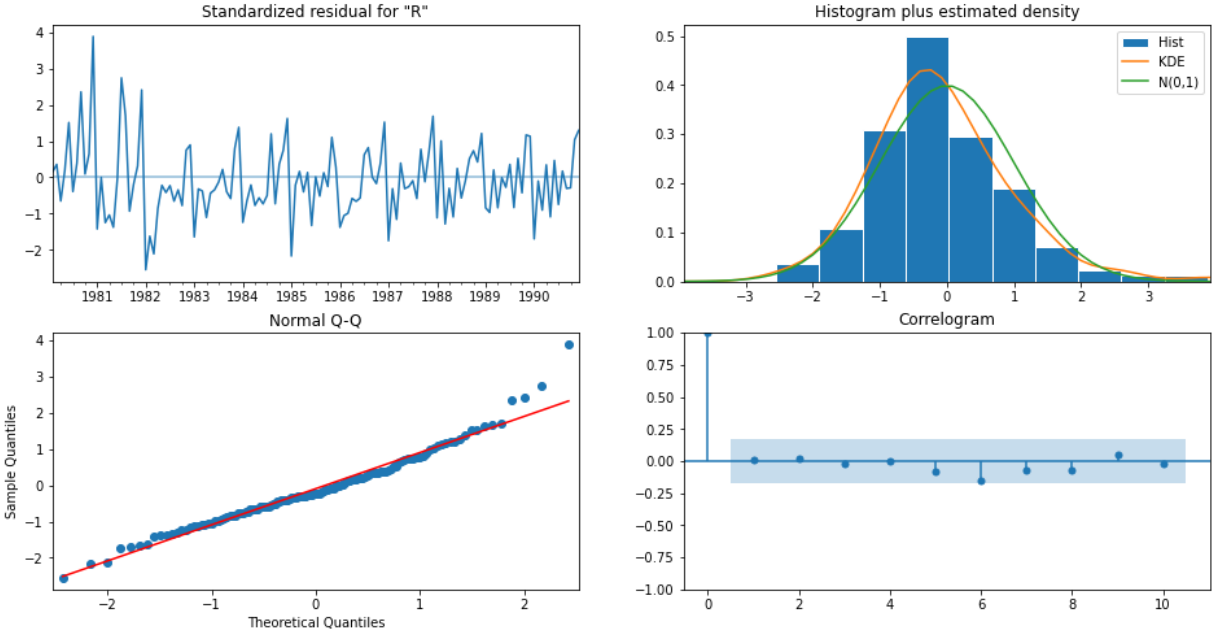

Fig. 63

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~36.

SARIMA Model:

- Different combinations of parameters (p, d, q) & (P, D, Q, F) defined for ARIMA model building are as shown below:

```
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

Fig. 64

- Let us fit the Training data set with the defined (p, d, q) & (P, D, Q, F) values and get the AIC scores. Sample of AIC scores obtained are as shown below:

```
SARIMA(0, 1, 0)x(0, 0, 0, 6) - AIC:1323.9657875279158
SARIMA(0, 1, 0)x(0, 0, 1, 6) - AIC:1264.4996261113856
```

Fig. 65

- AIC scores arranged in ascending order are shown below:

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	951.744297
59	(0, 1, 3)	(2, 0, 3, 6)	952.073632
251	(3, 1, 3)	(2, 0, 3, 6)	952.582102
191	(2, 1, 3)	(3, 0, 3, 6)	953.205651
123	(1, 1, 3)	(2, 0, 3, 6)	953.684951

Table. 27

- We can see that (2,1,3) & (2,0,3,6) combination of (p, d, q) & (P, D, Q, F) is giving the lowest AIC score for the SARIMA model. Let us use this combination to fit the Training set and Test set. Results are shown below:

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(2, 1, 3)x(2, 0, 3, 6)	Log Likelihood	-464.872
Date:	Mon, 17 Oct 2022	AIC	951.744
Time:	22:22:08	BIC	981.349
Sample:	01-01-1980 - 12-01-1990	HQIC	963.750
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5027	0.083	-6.081	0.000	-0.665	-0.341
ar.L2	-0.6628	0.084	-7.918	0.000	-0.827	-0.499
ma.L1	-0.3714	1622.483	-0.000	1.000	-3180.379	3179.636
ma.L2	0.2033	1019.901	0.000	1.000	-1998.766	1999.173
ma.L3	-0.8319	1349.770	-0.001	1.000	-2646.332	2644.668
ar.S.L6	-0.0838	0.049	-1.720	0.085	-0.179	0.012
ar.S.L12	0.8099	0.052	15.465	0.000	0.707	0.913
ma.S.L6	0.1702	0.248	0.686	0.493	-0.316	0.656
ma.S.L12	-0.5646	0.199	-2.834	0.005	-0.955	-0.174
ma.S.L18	0.1710	0.143	1.198	0.231	-0.109	0.451
sigma2	260.7795	4.23e+05	0.001	1.000	-8.29e+05	8.3e+05

Ljung-Box (L1) (Q):	0.72	Jarque-Bera (JB):	4.77
Prob(Q):	0.40	Prob(JB):	0.09
Heteroskedasticity (H):	0.54	Skew:	-0.36
Prob(H) (two-sided):	0.06	Kurtosis:	3.73

Fig. 66

- Plot diagnostics for the SARIMA model built are as shown below:

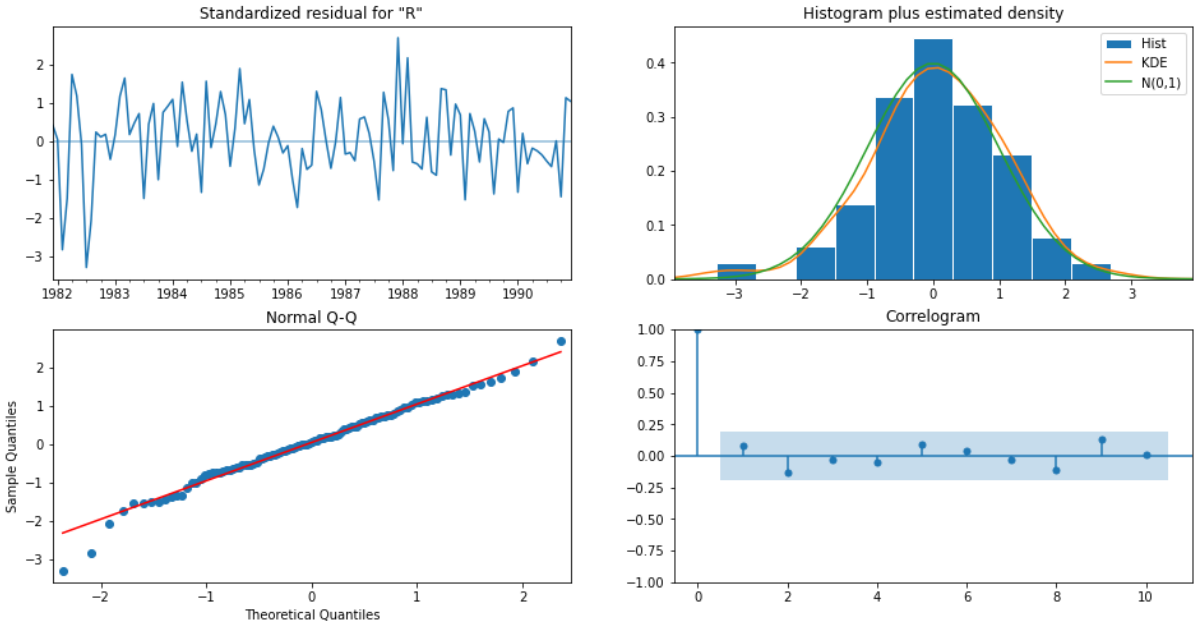

Fig. 67

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~27.

7. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

ACF plot:

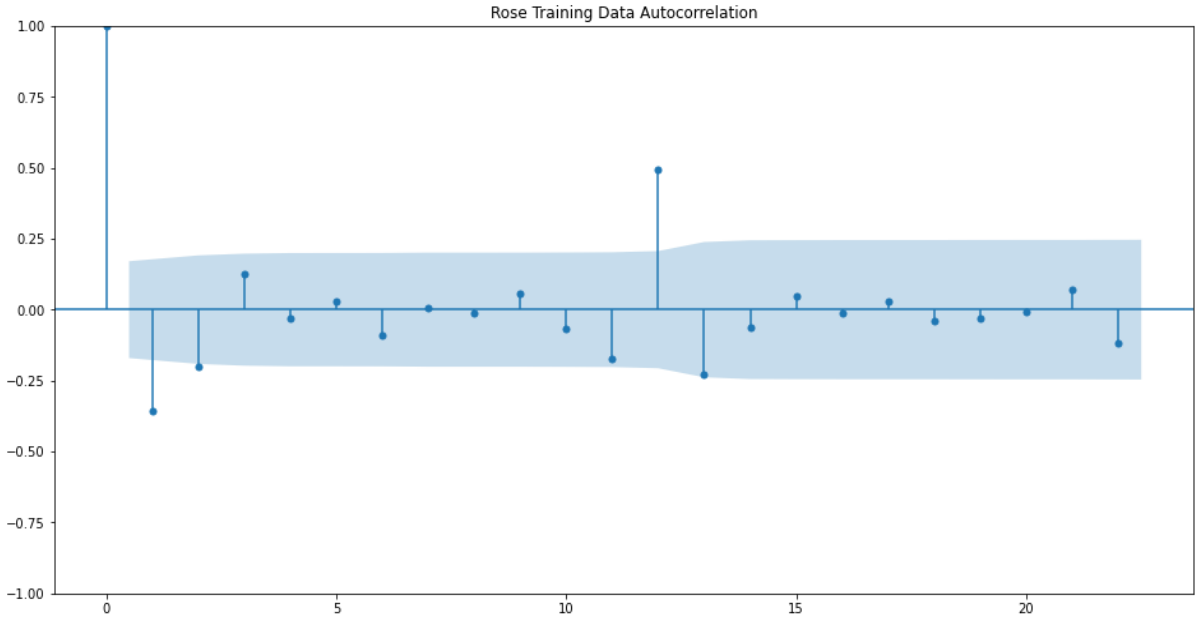

Fig. 68

- From above graph, q=3 and Q=0

PACF Plot:

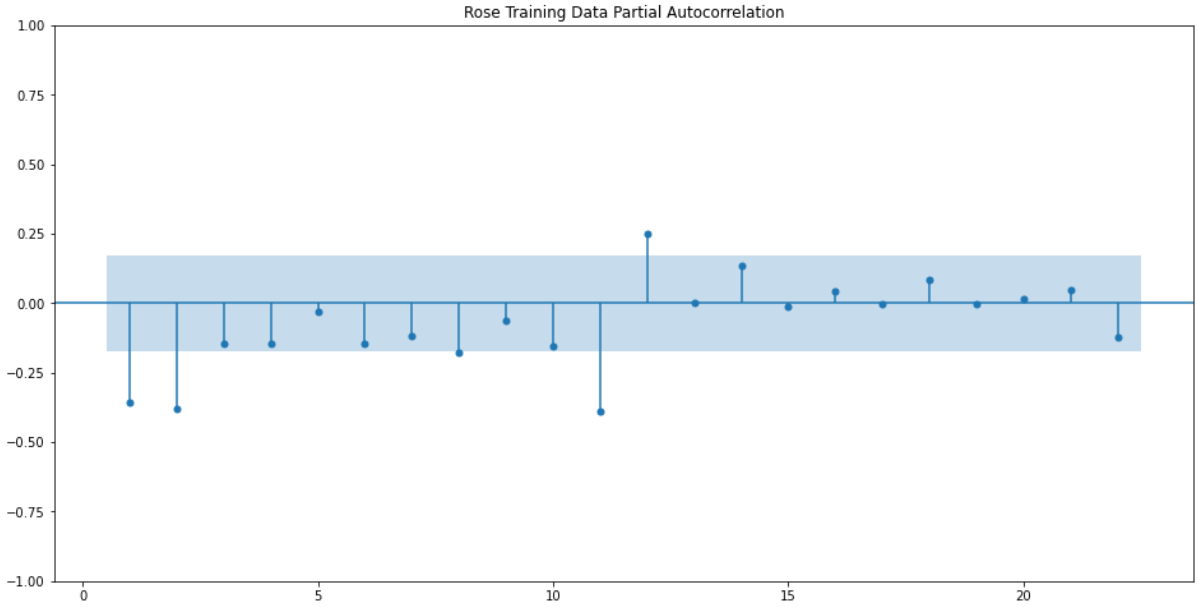

Fig. 69

- From graph, p=3 and P=0

- And we have already seen, 1st order difference of the data set is giving the stationary time series. So, d=1.

ARIMA Model:

- Let us fit the Training data set with the defined (p, d, q) = (3,1,2). Results are shown below:

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	ARIMA(3, 1, 3)	Log Likelihood	-632.331
Date:	Mon, 17 Oct 2022	AIC	1278.661
Time:	23:29:53	BIC	1298.788
Sample:	01-01-1980 - 12-01-1990	HQIC	1286.840
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.5877	0.088	-18.005	0.000	-1.761	-1.415
ar.L2	-0.6454	0.142	-4.545	0.000	-0.924	-0.367
ar.L3	0.1311	0.089	1.470	0.141	-0.044	0.306
ma.L1	0.9466	0.158	6.006	0.000	0.638	1.256
ma.L2	-0.7110	0.107	-6.651	0.000	-0.921	-0.501
ma.L3	-0.9082	0.152	-5.993	0.000	-1.205	-0.611
sigma2	874.9149	134.197	6.520	0.000	611.894	1137.936

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	31.52
Prob(Q):	0.94	Prob(JB):	0.00
Heteroskedasticity (H):	0.37	Skew:	0.72
Prob(H) (two-sided):	0.00	Kurtosis:	4.93

Fig. 70

- Plot diagnostics for the ARIMA model built are as shown below:

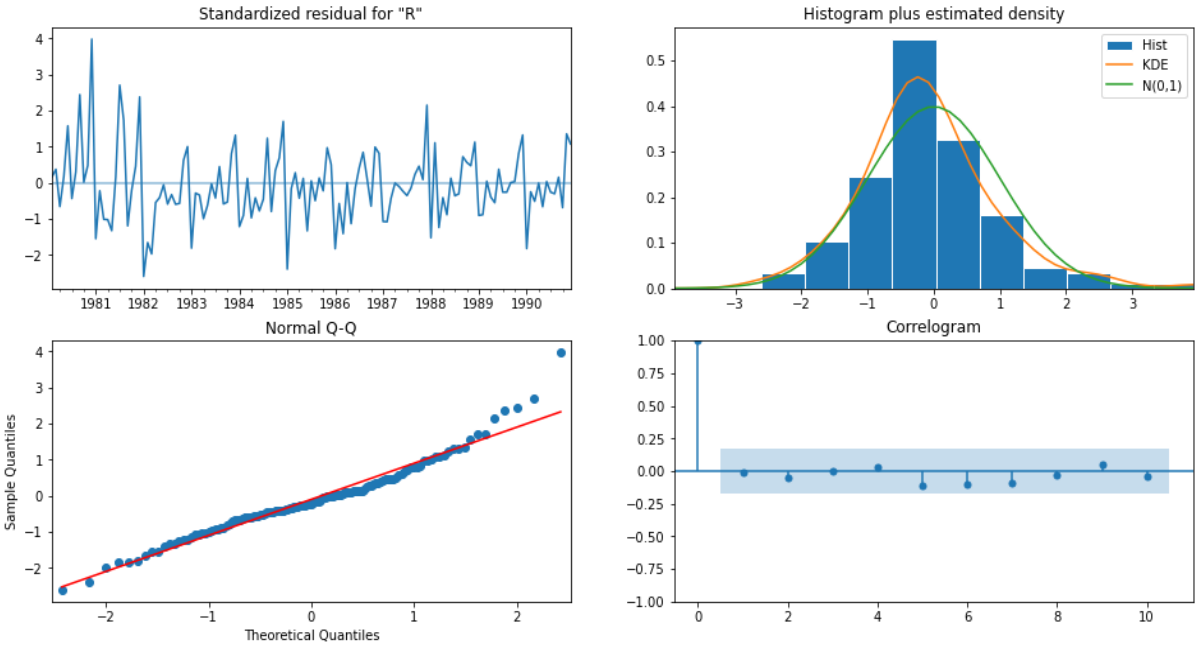

Fig. 71

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~36.

SARIMA Model:

- Let us fit the Training data set with the defined (p, d, q) = (3,1,3) & (P, D, Q, F) = (0,1,0,6). Results are shown below:

SARIMAX Results

Dep. Variable:	Rose	No. Observations:	132
Model:	SARIMAX(3, 1, 3)x(0, 1, [], 6)	Log Likelihood	-623.661
Date:	Mon, 17 Oct 2022	AIC	1261.322
Time:	23:38:23	BIC	1280.893
Sample:	01-01-1980 - 12-01-1990	HQIC	1269.271
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3438	0.093	-3.685	0.000	-0.527	-0.161
ar.L2	-0.7186	0.063	-11.454	0.000	-0.842	-0.596
ar.L3	-0.0084	0.084	-0.100	0.920	-0.172	0.155
ma.L1	-0.2915	329.540	-0.001	0.999	-646.178	645.595
ma.L2	0.2915	329.558	0.001	0.999	-645.630	646.213
ma.L3	-1.0000	0.131	-7.618	0.000	-1.257	-0.743
sigma2	1619.3876	0.114	1.42e+04	0.000	1619.164	1619.611

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	2.36
Prob(Q):	0.92	Prob(JB):	0.31
Heteroskedasticity (H):	0.39	Skew:	-0.28
Prob(H) (two-sided):	0.00	Kurtosis:	3.40

Fig. 72

- Plot diagnostics for the ARIMA model built are as shown below:

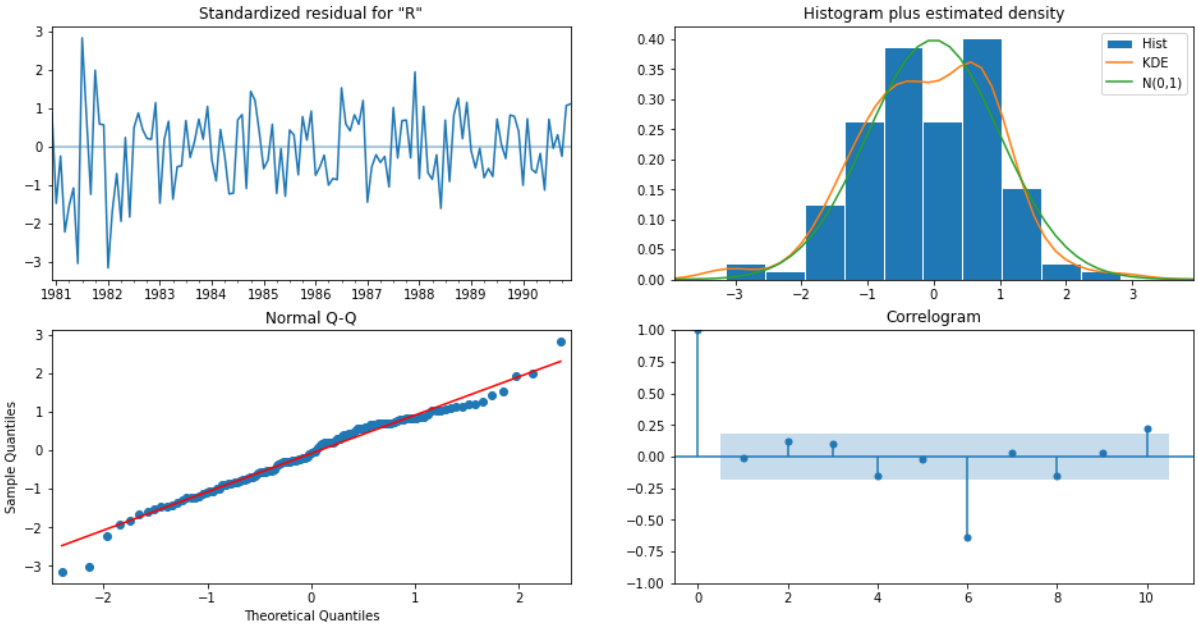

Fig. 73

- Let us predict on the Test set using the defined model.
- RMSE value for Test set is ~39.

8. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

Data frame with all the defined models and their test RMSE values:

	Test RMSE
RegressionOnTime	15.269416
NaiveModel	79.718824
SimpleAverageModel	53.460645
2pointTrailingMovingAverage	11.529409
4pointTrailingMovingAverage	14.450661
6pointTrailingMovingAverage	14.567606
8pointTrailingMovingAverage	14.807035
SESModel	36.796338
DESModel	15.269789
TESModel	14.263638
ARIMA_pdq_Model	36.815898
SARIMA_pdq_Model	27.124785
ARIMA_manual_Model	36.718313
SARIMA_manual_Model	39.533138

Table. 28

- 2-point moving average model is good along with 4,6,8-poinrt moving average model and TES model.

9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

- Let us predict the 12 months forecast by using the TES Model.
- Let us fit the whole data frame first and best parameters are as shown below:

```
{'smoothing_level': 0.09688703220086504,
 'smoothing_trend': 1.153463764049725e-05,
 'smoothing_seasonal': 8.610887784810513e-05,
 'damping_trend': nan,
 'initial_level': 145.50189021278803,
 'initial_trend': -0.5373158948766142,
 'initial_seasons': array([-28.07666039, -17.21945901,  -9.119496  , -15.77603716,
       -11.86135785,  -5.83432707,   5.30507007,   5.31697715,
         2.67072552,   1.9040477 ,  17.03158686,  55.90270161]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Fig. 74

- Next 12 months forecast sales' values predicted by TES model are as shown below:

```
1995-08-01    50.018600
1995-09-01    46.835018
1995-10-01    45.531078
1995-11-01    60.121312
1995-12-01    98.455017
1996-01-01    13.938433
1996-02-01    24.258307
1996-03-01    31.820954
1996-04-01    24.627111
1996-05-01    28.004465
1996-06-01    33.494194
1996-07-01    44.095214
Freq: MS, dtype: float64
```

Fig. 75

- Plot showing whole data frame and 12 months forecast is as shown below:

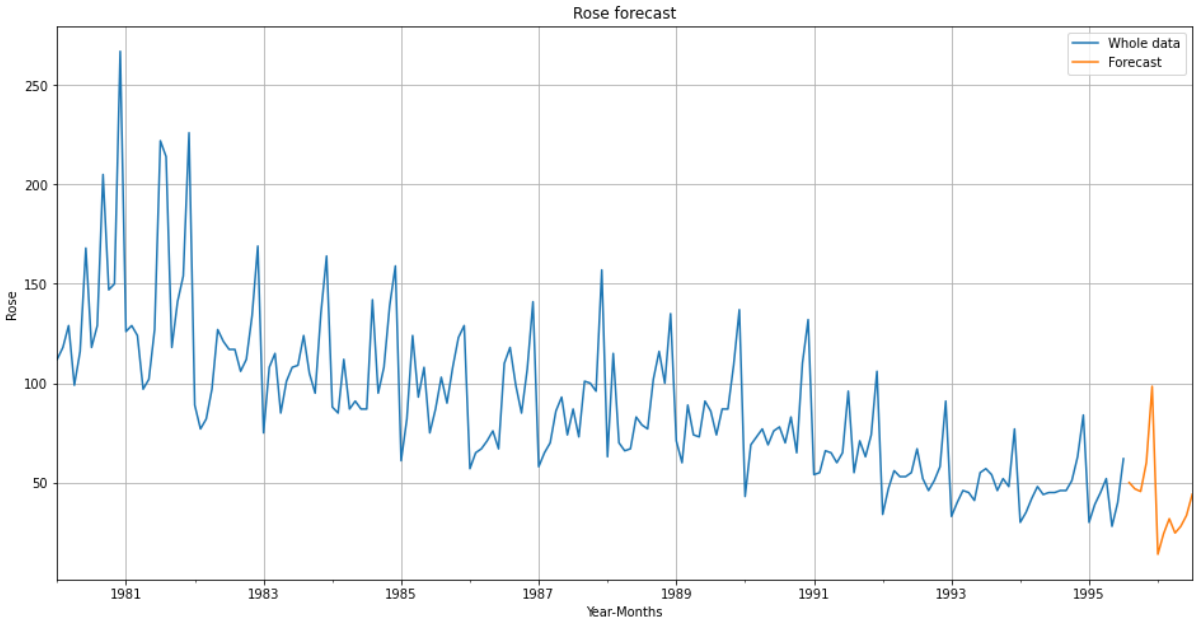

Fig. 76

10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- Rose wines also winter season demanded.
- Rose wines are losing their demand in the market because Times series showing clear decreasing trend over the years.
- There is a rapid decrease in sales from 1982 to 1983. Company should investigate seriously for the reasons behind this serious rapid fall. This investigation will help in analysing the decrease trend of sales.
- Forecast of 12 months is showing increase in sales which is a good thing. Company should capitalize on this.

THE END